UNIVERSITY
OF FLORIDA
LIBRARIES

1855

SCIENCE ROOM

Digitized by the Internet Archive
in 2010 with funding from
Lyrasis Members and Sloan Foundation

http://www.archive.org/details/measurementsofhu00gree

MEASUREMENTS OF HUMAN BEHAVIOR

Left:

Lewis M. Terman

(Photo by Hans Roth)

Above:

Edward L. Thorndike

(Photo by Harris & Ewing)

Right:

L. L. Thurstone

(Photo by Underwood & Underwood)

MEASUREMENTS OF HUMAN BEHAVIOR

EDWARD B. GREENE
Assistant Professor of Psychology
University of Michigan

THE ODYSSEY PRESS
New York

COPYRIGHT, 1941
BY THE ODYSSEY PRESS, INC.
ALL RIGHTS RESERVED

PRINTED IN THE UNITED STATES
SECOND PRINTING

THIS BOOK IS
AFFECTIONATELY DEDICATED
TO
MY MOTHER AND FATHER

147360

PREFACE

In recent years there have been two developments, each of which has contributed greatly to individual human adjustment. One is the growth of experimental psychology into a fairly mature biological science, the result of the efforts of a few hundred research workers scattered all over the world. The other is the rapid expansion of individual guidance in educational, industrial, and clinical centers, the result of the efforts of many thousands of persons. There are in the United States today approximately 22,000 members of the American Vocational Association, and probably as many more in employment and clinical fields, who are devoting most of their time to the problems of analyzing individual abilities and needs, and of helping others to plan their lives more satisfactorily. In addition, there are approximately 650,000 elementary school teachers, 260,000 high school teachers, and 90,000 college teachers in the United States, practically all of whom spend considerable time in appraising student achievement by means of formal or informal tests.

Although both scientific research workers and those engaged in practical guidance are making observations of human behavior, an enormous gap usually exists between them. The scientist analyzes patterns of behavior, whereas the practical worker observes differences between persons which seem to be of practical, moral, or aesthetic importance. All too often the scientist berates the practical worker for placing values upon behavior without knowing very much about its nature and causes. Conversely, the practical worker accuses the research worker of being uninterested in the important problems of life adjustments. Between these two extremes of both purpose and method there lies a middle ground. Here attempts are made to combine the best qualitative analyses with the most appropriate quantitative appraisals of individual differences. One result of the work carried on during the past forty years has been the extensive development of instruments for appraising human abilities, traits, interests, and attitudes. This development has in turn led to the introduction into college curricula of courses which deal with standard appraisals of behavior in human beings. This book is a response to the desire of teachers for a text which will present a more comprehensive treatment of behavioral measurement techniques and procedures than is available in any single volume hitherto published. It has been planned to serve the needs of those students in the second, third, and fourth years of college who are preparing for work in medicine, journalism, education, industrial relations, social work, or applied psychology—students who, as a rule, cannot take time to read many original sources, even when they are available.

The book is arranged in three parts. All of Part I, which deals with basic considerations in measurement, must be understood by those who wish to study the rest of the book. Most of the chapters in Part II, which deals with typical measuring instruments and their uses, and in Part III, which deals with persistent problems in measurement, may be read separately without serious loss of context. In a short course, such chapters should be selected as seem best adapted to satisfy the interests and needs of the particular students enrolled.

Throughout the text emphasis is placed on a clear presentation of facts about testing instruments, and their interpretation. The differences between adequate and inadequate practices have been pointed out so frequently that the establishment of criteria of accuracy has been selected as one of the major themes of the book. Although norms for various age or occupational groups are frequently given, the reader is referred to other references for systematic discussions of the uses of tests in industry, schools, and clinics. Controversial issues are often outlined but are seldom discussed at length. Many needed researches are described, however, in the hope that some readers will be stimulated either to take part in investigations or to become alert to new contributions in this rapidly growing field.

Owing to the wealth of material from which he could choose, the author has in many instances found the selection of illustrative items very difficult. In this connection it should be pointed out that European work has been infrequently mentioned, and that no attempt has been made to list all of the important tests in this country; there are compendia which do this job annually. All the important types of tests and appraisals have, however, been represented, among them being a number that are not usually found in texts on mental measurement. Included in this latter category are appraisals of artistic ability, interests, and attitudes. These are considered to be essential facts of behavior, and they may be measured by methods similar to those used for appraisals of cognitive phenomena. Such appraisals are valuable in that they add important aspects to one's knowledge of a person, aspects which, fortunately, are being emphasized much more today than ever before.

Among the distinctive features in this book is the discussion given to methods and to the results of factorial analyses. Without attempting to present the mathematics of factorial solutions, which is generally reserved for the course in statistics, the text covers the logic, which is very important in analytical thinking, and the results, which are significant in test construction and interpretation.

Other features, which will be particularly useful for students, are the combination of a glossary and a subject index, and the use of a complete bibliography to serve as an author index.

Acknowledgments to authors and publishers of quoted works are made at appropriate places in the text. Thanks and appreciation are due to my wife, Katharine B. Greene, who composed the chapter on "Tests of Early Childhood" and criticized the whole manuscript constructively. Thanks for valuable suggestions are also extended to the following who have read one or more of the chapters in manuscript: Professors Henry F. Adams, Martha G. Colby, C. W. Crannell, Paul S. Dwyer, George Meyer, W. C. Olson, Burton D. Thuma, L. L. Thurstone, W. B. Pillsbury, John F. Shepard, and Edna Gordon. Professor E. R. Henry and Professor Philip L. Harriman have read the entire manuscript and made many practical suggestions.

Invaluable aid has also been given by the following students, who pointed out places in the text where more adequate explanations were needed: Edward Varnum, Robert Kleemeier, William Gilbert, Alastair Craig, Robert E. Fryer, J. McCormak, and Virginia H. Johnson.

Finally, thanks are gratefully given to the editorial and production staffs of The Odyssey Press, whose judgment has made the book more useful than it could have been without their help.

EDWARD B. GREENE

Ann Arbor, Michigan
April, 1941

CONTENTS

Part I

BASIC CONSIDERATIONS

CHAPTER PAGE

1. **Introduction** 3
   - I. What Is Behavior? 3
   - II. What Is Measurement? 4
     - How Are Scales Developed? 5
   - III. The Application of Standard Scales 7
     - A. Experimental Use 7
     - B. Individual Appraisals 7
     - C. Detecting Group Tendencies 8
     - D. Agencies Which Use Tests 9
   - IV. The Requirements of a Good Examiner 12
   - V. Limitations of Testing Techniques 13

2. **Varieties of Appraisals** 15
   - I. Well-Controlled Situations 15
     - A. Physiological Measures 16
     - B. Psychophysical Measures 17
     - C. Single Variable Tests 17
   - II. Partly Controlled Situations 20
     - A. Individual and Group Testing 20
     - B. Measures of Skill 21
     - C. Measures of Retention 29
     - D. Measures of Learning 32
   - III. Slightly Controlled Situations 34
     - A. Time-Sampling 34
     - B. Ratings 35
     - C. Logs 37
     - D. Interviews 37
     - E. Case Histories and Biographies 38

CHAPTER PAGE
3. **The Interpretation of Scores** 39
    I. The Limits of a Score 40
    II. Frequency or Distribution Tables 40
    III. Histograms and Frequency Curves 41
    IV. Normal Frequency Curve 44
    V. Ogive Curve 45
    VI. Ranks, Centiles, Quartiles, Deciles, and Letter Grades 47
    VII. Dimensions of a Group of Scores 50
        A. Central Tendencies 50
        B. Dispersions 55
    VIII. Uses of Measures of Dispersion 58
        A. Group Comparisons 58
        B. Individual Comparisons 59
        C. Standard Score versus Centiles 61
        D. Profiles 61
        E. To Combine a Person's Score on Several Tests 62
        F. Weighted Scores 63

4. **Measures of Relationship** 65
    I. Predictions of Probable Scores from Known Scores 65
        A. Scatter Diagrams 66
        B. Regression Lines 68
        C. Slope of Regression Lines; Correlation Coefficients 69
        D. Product-Moment Correlation Coefficients 71
        E. Rank-Order Correlation Coefficients 72
        F. Individual Predictions 73
        G. Standard-Error-of-Estimate 74
        H. Standard-Error-of-a-Score 75
    II. Applications of Correlation Techniques 76
        A. Related Persons 77
        B. Related Scores 77
        C. Related Items 77
    III. Errors in Applying Correlation Coefficients 77
        A. Errors Which Usually Lower Correlations below Their True Value 78
        B. Errors Which Raise Correlations Spuriously 81

CHAPTER PAGE
IV. How Large Must Correlations Be for the Purpose of Predictions? 82
A. Probable Error of a Correlation Coefficient 83
B. Coefficient of Alienation (k) 84
V. Theoretical Interpretations of Correlations 86
A. Identical Elements 87
B. Covariation 89

5. **Measuring Instruments** 90
I. Types of Tests 90
A. Definitions 90
B. Formal and Informal Tests 91
C. Achievement Tests 91
D. Aptitude Tests 92
E. Psychological Tests 94
F. Altitude, Speed, and Breadth Tests 94
II. Characteristics of a Good Instrument 97
A. Validity 97
B. Reliability 100
C. Equivalence of Steps 105
D. Practical Aspects 106

6. **Construction and Evaluation of Test Items** 109
I. Types of Test Items 109
II. Merits of Various Types 110
III. Relative Time Allowances 115
IV. The Construction of Informal Appraisals 116
V. Rules for Item Construction 117
VI. Statistical Analysis of Items 121
A. Detection of Ineffective Items 121
B. Chance Success 123
VII. Efficient Scoring Devices 131
A. Stencil Scoring 131
B. Automatic Scoring Devices 132
C. Machine Scoring 134

7. **Factorial Analyses** 135
I. Assumptions 135
II. Factors 137
III. Spearman's Contribution 138
A. The General Factor (g) 139
B. Group Factors 139

CHAPTER PAGE
IV. Thurstone's Method 141
    A. Analyses of a Score 141
    B. Analyses of a Group of Scores on a Test 143
    C. Common Factors 145
    D. Chance and Specific Factors 146
V. Limitations and Advantages 146
VI. Typical Results of Factor Analysis 148

Part II

INSTRUMENTS AND RESULTS

8. Tests of Early Childhood 151
I. Introduction 151
II. Description of Scales 153
    A. Gesell and Thompson (1938) Norms of Early Growth, 4 to 56 Weeks 153
    B. Gesell's Development Schedules, 0 to 60 Months 159
    C. Bühler and Hetzer (1935), Birth to 6 Years 160
    D. Merrill-Palmer Scale, 18 to 63 Months (Stutsman, 1931) 162
    E. Minnesota Preschool Tests, 1½ to 6 Years (Goodenough, 1932) 165
    F. Goodenough Measurement of Intelligence by Drawing, 3½ to 13½ Years 167
    G. The Stanford-Binet and the Kuhlmann-Binet Tests 169
    H. The California Preschool Schedule 169
    I. Measures of Reading Readiness 170
    J. Summary of Description of Scales 175
III. Time Sampling 176
IV. Some Results of Measurement 177
    A. The Usual Order of Development 177
    B. Individual Profiles 181
    C. Correlational Studies of Growth 182
✓ **9. Measures of Achievement** 187
I. Cardinal Objectives 187
II. Measurement of Language 189
    A. Appraisals of Expression 191
    B. Appraisals of Comprehension 199

CHAPTER PAGE
III. Measurement of Mathematical Ability 227
A. Arithmetic Tests 227
B. Algebra, Geometry, and Trigonometry Tests 229
IV. Special Studies 233
A. Physical Sciences 234
B. Social Sciences 236
C. Foreign Languages 240
D. Business Achievement 241
V. Study Habits 246
VI. Indices of Accomplishment 248
Accomplishment Quotients 250
VII. Correlational Analyses 251
VIII. Applications 256
A. Scholastic Predictions 256
B. Optimum Age for Instruction 259
C. Vocational Predictions 261

10. Binet-Type Scales 264
I. Characteristics 264
II. Early Scales 265
III. Terman's Revisions 270
IV. Kuhlmann's Revisions 273
V. Other Revisions 275
VI. Construction of Binet-Type Tests 277
A. Selection of Items 277
B. Scaling of Items 282
VII. Interpretation of Mental Ages 285
A. Inequality of Steps 285
B. Changes in Meaning of Mental Ages above Adult Level 285
C. Processes Measured at Different Ages 287
D. Components of Equal Mental Ages 287
VIII. Interpretation of Intelligence Quotients 288
A. Similarity of Average IQ's 289
B. Similarity of Standard Deviations of IQ's 289
C. Calculation for Adults 291
D. Constancy of IQ's 292
IX. Substitutes for Mental Ages and Intelligence Quotients 295
X. Principal Uses 297

CHAPTER PAGE
11. **Group Intelligence Tests** 298
  I. Introduction 298
  II. The United States Army Tests 305
    A. The Alpha Test 308
    B. The Beta Test 309
  III. Other Group Tests 319
    A. The California Tests of Mental Maturity 323
    B. McCall, Herring, Loftus Comprehensive Battery 323
    C. Thorndike's CAVD Scales 325
    D. Kuhlmann-Anderson (1927) Intelligence Tests 328
    E. Thurstone's Test of Primary Traits 329
    F. The Bellevue Scale 334
  IV. Correlational Analysis 337
    A. Unitary Traits 337
    B. The Effects of Difficulty on Obtained Factors 340
    C. Discussion of Analysis 341
12. **Performance, Mechanical, and Motor Tests** 343
  I. Introduction 343
  II. Performance Tests 344
    A. Pintner-Paterson Scale 344
    B. Arthur Performance Scale 350
    C. Other Scales 352
  III. Mechanical Ability Tests 354
    A. Paper and Pencil Tests 354
    B. Assembly Tests 358
  IV. Motor Co-ordination Tests 363
    A. Reaction Time 364
    B. Agility and Strength 367
    C. Dexterity 371
    D. Steadiness 373
    E. Motor Rhythms Tests 374
    F. Batteries of Motor Tests 376
  V. Correlational Analyses 378
13. **Measurement in Fine Arts: Design** 385
  I. The Nature of Art 385
  II. The Nature of Appreciation 386
    A. Tests of Preferences for Pictures 393

CHAPTER PAGE
B. Responses to Lines 397
C. Responses to Color 399
D. Discussion of Tests 402
III. The Nature of Artistic Composition 403
A. Development of Drawing Ability 405
B. Tests of Composition and Representation 406
IV. Intelligence and Art Tests 409
V. Analytical Studies of Artistic Ability 411

14. Measurement in Fine Arts: Literature and Music 415
I. Measurement in Literature 415
A. Tests of Discrimination 415
B. Tests of Appreciation 419
C. Tests of Information 423
D. Appraisals of Compositions 424
E. Analysis of Motives 425
II. Measurement in Music 425
A. Discrimination 426
B. Appreciation of Musical Stimuli 428
C. Accomplishment 430
D. Correlational Analyses 433
E. Applications 435

15. Academic and Vocational Interests 439
I. Definitions 439
II. Typical Studies 442
A. Case Histories 442
B. Reasons for Choices 445
C. Reasons for Job Satisfaction 449
D. Inventories of Preferences 457
III. Practical Results 473
A. Permanence of Interest 473
B. Prediction of Scholastic Success 477
C. Vocational Predictions from Interest Scores 481
IV. Future Research 483

16. Appraisals of Attitudes 485
I. Introduction 485
II. Typical Scales 486
A. Watson's Test of Public Opinion 486
B. Thurstone and Chave's Attitudes Scales 491
C. Murphy and Likert's Study of Public Opinion 492

CHAPTER PAGE
D. Attitude toward the Negro 496
E. Reasons for Acts 498
F. Allport and Vernon's Study of Values 501
G. Factorial Analyses of Attitudes 506
III. Practical Results 510
A. Modification of Attitudes 510
B. The Relation of Information to Attitudes 515
C. Comparison of Various Techniques 516
D. Group Norms 519
IV. Needed Research 520

17. Modes of Adjustment: Free Association and Inventories 521
I. Introduction 521
II. Free Association 523
A. The Rorschach Test 524
B. Picture Interpretation and Composition 533
C. Word Association Tests 536
III. Questionnaires 538
A. Psychiatric Classification 540
B. Psychological Theory 540

18. Modes of Adjustment: Conduct Measures, Observations, and Laboratory Methods 549
I. Introduction 549
II. Conduct Measures 549
A. Measures of Deception, Hartshorne and May (1928) 550
B. Measures of Co-operation, Hartshorne and May (1929), Chapter III 552
C. Measures of Persistence and Inhibition 555
D. Tests of Moral Knowledge, Hartshorne and May (1930), Chapter IX 556
E. Tests of Opinion, Hartshorne and May (1930) 558
F. Measures of Integration 559
III. Records of Observations 562
A. Continuous Records 562
B. Time-Sampling 569
C. Records of Interviews and Test Situations 570
IV. Laboratory Methods 570

CHAPTER PAGE

19. Modes of Adjustment: Evaluation of Methods and Results 579
- I. Comparison of Methods 579
  - A. Economy 579
  - B. Reliability 580
  - C. Validity 581
- II. Correlational Analyses 583
  - A. Results from Thurstone's Method 584
  - B. Spearman's Method 589
  - C. Conclusions 593
- III. Evaluation of Items 594
  - A. Elimination of Ambiguities 595
  - B. Emotional Reactions to Items 596

Part III

PERSISTENT PROBLEMS

20. Effects of Practice on Test Scores 601
- I. Significance of Practice 601
- II. Conditions for Measurement of Practice 602
  - A. Equal Amounts of Practice 602
  - B. Equal Motivation 603
  - C. Constant Behavior Patterns 603
  - D. Similar Types of Scores 604
  - E. Equal Units of Measurement 606
  - F. True Zeros of Development 607
- III. A Universal Practice Curve 607
- IV. Effects of Practice on Growth Curves 612
- V. Measured Results of Practice on Tests 613
  - A. Changes in Mean Scores 614
  - B. Changes in Group Dispersion 623
  - C. Correlation of Trials. Prediction 624
  - D. Relations between Measured Changes 629
  - E. Factorial Analyses of Practice 630
- VI. Avoidance of Practice Effects 633
- VII. Measured Results of Forgetting 634

21. Measurement of Growth and Senescence 638
- I. The Nature of Growth 638
- II. Conditions Essential for Growth Studies 641

CHAPTER PAGE
III. Typical Results of Growth Studies 643
A. Anatomical Structures 643
B. Motor Co-ordination and Spatial Thinking 647
C. Intellectual Growth 650
D. Character Traits 658
IV. Dispersion of Scores 659
V. Relationships among Various Growth Patterns 663
A. Correlational Studies 663
B. Individual Growth Curves 665
VI. Theoretical Considerations 668
A. Simple or G Curves 668
B. Cycles: Successive Combinations 671
C. Simultaneous Combinations 672
D. Suggested Scales 672

22. Standard Deviation or Absolute Scaling 675
I. The Need 675
II. The Nature of SD Scaling 677
III. Calculation of SD Scale Values 678
SD Scaling of Total Scores 683
IV. Limitations of the Method 684
V. Some Results of SD Scaling 684
A. Selection of Items 685
B. Age Group Dispersions 685
C. True Zeros 686
D. Mean Age Curves 687
E. Natural Laws 690

23. The Evaluation of Judgments 691
I. The Problem 691
II. Paired-Comparisons Method 692
III. Rank-Order Method 694
IV. Rating-Scale Methods 699
A. Forms of Rating Scales 699
B. Advantages of Various Forms 701
C. Relative Intangibility of Items 703
D. Number of Desirable Steps 704
E. Halos and Logical Errors 705
F. Rules for Raters 707
G. Combinations of Ratings 707
H. Validity of Ratings 708

CHAPTER PAGE
24. Measurement of Native Differences 713
I. Selective Breeding 713
A. Relatives of Superior Persons 714
B. Relatives of Inferior Persons 715
C. Correlational Studies 715
D. Experimental Studies 717
II. Environmental Changes 719
A. Changes in Home Environment 719
B. Special Training 724

Bibliography 733
Subject Index 765

Part I

BASIC CONSIDERATIONS

1

INTRODUCTION

This chapter answers concisely each of five important questions about the measurement of human behavior:

1. What is implied in the term *behavior?*
2. What is measurement?
3. What are the main uses of measures of behavior?
4. What requirements should a good examiner meet?
5. What are the limitations of measurement?

Other chapters in this book give more detailed answers to these questions.

I. WHAT IS BEHAVIOR?

The term *behavior* refers to any series of acts of an individual which occur in a particular place during a particular time. The individual may be an object, a person, or a hypothetical entity, such as an electron. The acts may be thought of as purely physical or as involving mental phenomena. No attempt will be made here to lay down rigid technical discriminations between physical and mental phenomena. In general, physical phenomena refer to forces, movements, and qualities of chemical elements; or to combinations of elements, either animate or inanimate. Under mental phenomena are grouped acts of living organisms which are difficult to describe entirely in physical terms. These acts include wishes, memories of experiences, and beliefs, most of which involve verbal symbols, and all of which seem to be dependent upon the activity of nervous tissue. Many investigators believe that mental and physical acts can be explained by the same natural laws; hence, that there is no difference between them. Certainly there is no sharp line separating them. For convenience, mental acts may be considered a subclass of physical phenomena which, at present, must be studied indirectly. The indirect approach is usually necessary, for the brain elements are too small and too easily destroyed to be observed and measured directly. The nature of mental acts and mental organization is discussed throughout the book, and particularly in Chapters 2, 7, 19, and 24.

II. WHAT IS MEASUREMENT?

Broadly speaking, measurement is a term used to represent any sort of comparison reported in some numerical fashion. According to this definition, all measurements involve two somewhat independent processes: first, a judgment of fact; and, second, a mathematical procedure, called scaling, which attaches a number, called a score, to the results of the comparison. Practically, there are two types of comparisons:

1. Qualitative comparisons. Here, two or more persons or objects are compared to determine whether they have the same qualities. Qualitative judgments of a rough sort would be used in the examination of two persons to see whether they could both hear, or could both solve arithmetic problems.
2. Quantitative comparisons. After qualitative similari-

ties have been established, comparisons of amount can be made. Judgments of this sort are illustrated by estimating which of two persons has the more acute hearing, or which has more arithmetical ability.

From these two types of comparisons, convenient units of amount can sometimes be specified with great precision. Chapter 2 and Part II describe the most frequently used units of comparisons for appraising human abilities.

How Are Scales Developed?

Many of the measurement techniques used by psychologists were partly developed by persons working in physical sciences and were then adapted to the appraisal of activities of persons or animals. This fact makes it possible for persons well trained in measurement techniques to understand one another's methods fairly well, even though they work in different fields. All scales of measurement have been developed along similar lines.

A good measuring scale is the result of many years of hard work. The procedure is well illustrated by the development of measures of the hearing of a boy who is suspected of being somewhat deaf. One of the crudest measures is made by observing him to see whether he responds to ordinary sounds by turning his head or trying to see what made the sounds. One cannot definitely establish deafness from such observation, since the child may be feeble-minded or unco-operative and therefore may not respond normally.

For a more careful diagnosis, the youngster may be taken into a quiet room and asked to tell whether or not he hears a watch ticking. The watch may be held at various distances from one ear while his other ear is covered. This method of ascertaining deafness is better than the first, but it is neither complete nor accurate. The youngster may be deaf to certain tones only, or he may say that he hears the watch when he does not.

A further refinement in the measurement of hearing can be made with an audiometer. This is a machine which speaks into one ear at a time. The child must report the numbers which he hears. The best audiometers give a wide sample of vowel and consonant sounds at various intensities. A record may be obtained which shows all of the child's answers. From these answers, a measure of his range of hearing for pitch, intensity, and vowels and consonants may be obtained. Because there may be some chance successes, however, the test may not be entirely accurate and complete. Also, if the child is very young or mentally retarded, or if he has a speech defect, he may not be able to report correctly the sounds he hears. Hence, this measurement of hearing is limited by the subject's ability to report sounds. A still more precise way of measuring hearing does not require the reproduction of sounds, but presents two sounds following which the child is asked to indicate by a simple movement whether they are the same or different.

There are still some unknown factors present in this situation. For instance, we can never be sure that a child has done his very best. If the test situation is somewhat strange and terrifying to him, he may wish to co-operate but fail to do so. If the boy is suffering from a head cold or fatigue, he may not be able to do his best. In spite of all precautions, therefore, one must admit that the best appraisals are only approximations of actual facts rather than true measures. This same admission must be made for all measurement in any science.

From this account of the development of measurement of hearing it appears that there existed at first a rough concept of not being able to hear. After much experimentation, this concept was discarded. In its place, three more precise concepts which can be appraised numerically are now used. These are pitch discrimination, intensity discrimination, and discrimination of spoken sounds which are complex combinations of various pitches, intensities, and rhythms. The construction of instruments for measuring each sort of hearing phenomenon requires much mechanical skill. Refinements that insure still greater precision are being discovered from time to time. The construction of scales is discussed in Chapters 5, 22, and 23 in more detail.

III. THE APPLICATION OF STANDARD SCALES

Mental scales have three main uses, all of which lead to refinements in the prediction of human development. One is for experimentation, another for individual appraisal, and a third for detecting group tendencies.

A. Experimental Use

Tests are often used experimentally to find out, for instance, whether one method of learning is better than another, whether insight is used to solve a particular problem, whether two persons have the same form of growth curve, whether individual output is better under one type of lighting than under another, or whether one test predicts success in learning to read better than another. Such experimental applications of tests are necessary before the tests can be used for predictions of success.

B. Individual Appraisals

An important individual use of mental measurement lies in showing one what his strongest and weakest points are. A person usually has a rough idea of his own abilities and interests, but he finds in test results a more accurate indication of where he stands in a group. The results of individual analysis are shown in all the chapters in Part II.

Tests are also important in situations where a decision is to be made about a person, such as whether he shall take a certain course of study, whether he is to be offered a position, whether he shall be committed to an institution, or whether he shall marry *that* girl. The scores of an individual may be compared with the scores of others whose development has been studied over a long period of time. From such comparisons predictions of success can be made with a known degree of probability. Techniques for making predictions are discussed in Chapters 3, 4, and 21.

C. Detecting Group Tendencies

Very often test results which are not accurate enough to be used for individual predictions are valuable in showing group tendencies. This statement is true, for although the behavior of many persons in a group varies from time to time, the scores of the group often remain constant for long periods of time. In a large group, individual variations tend to occur in such a way as to balance one another. On one day some persons in a group may be poorly motivated, while others may be highly stimulated; on another day the reverse may be true. The group results for the two days will be the same, although some individual scores may vary enormously.

The use of tests to evaluate group tendencies is valuable in many situations, among them the selection of employees and the determination of the best methods of instruction. From the results of tests the level of efficiency of employees can usually be raised, and the best method of instruction for the average person can be ascertained. It is well, however, to avoid generalizing too freely. There are usually individuals who do not pass the tests very well, but who would make excellent employees. A high test score alone seldom indicates employability, since factors not at present represented in tests are often important in success. Likewise, the method of learning which is best for the average students in a group is likely to be a poor method for some of the others. Generalizations from group tendencies have probably been made too frequently and too emphatically in many fields. The best predictions of growth and practice come from a careful analysis of individual development.

There are no sharp lines among tests for individual, group, and experimental uses. Carefully recorded measurements may be, and often are, used to support a theory, even though they were made for individual use. For example, Strong's (1931) Vocational Interest Blank may be used to find out what a person's dominant interests are, or the same blank may be used experimentally to see whether interests change with age or experiences. Likewise, a test of mechanical information may be used practically to select persons for a mechanical trade, or experimentally to find out which of two learning methods is most efficacious. A test of visual acuity may be used to fit a person with glasses or to discover whether or not conscious visual cues are used in a particular problem.

In the final analysis, the major objectives of all applications of tests are nearly the same – namely, the accurate knowledge of present status, the prediction of development, and hence, a program of human progress. More accurate prediction leads to better planning, for when one knows with considerable certainty that he will progress with greater interest and effectiveness under one procedure than under another, it is his responsibility to strive for the better procedure. Likewise, when one knows that the chances that a particular person will develop criminal tendencies are very high in a particular situation, effort should be made to try to avoid his meeting that situation. Predictions of behavior are at present much better than mere chance, and in the future they will be still more accurate.

D. Agencies Which Use Tests

Four kinds of agencies – educational, industrial, clinical, and civic – regularly apply standard measures of behavior.

Educators use tests either for individual diagnosis and promotion, or for appraisals of a method of instruction or of an instructor. The diagnostic use of tests is often of great benefit to the student if his failings are recognized and a remedial course is made available for him. The greatest benefits are seen when vocational and educational counseling are combined with teaching, for counseling involves a plan of individual development. In order to make such a plan, the counselor needs the detailed information yielded by the most accurate tests.

The use of standard tests has in some cases led to undesirable results. Teachers have felt driven to prepare their pupils to meet certain test requirements rather than to develop in the students a mastery of skills in a reasonable sequence. Many progressive educators consider that this drive to insure that pupils meet certain requirements is a serious menace to good teaching. Surveys by means of tests are of considerable value both to administrators and to teachers when the goals of instruction are not sacrificed to coaching or cramming procedures.

Industrial agencies have used tests principally in the selection or promotion of employees. Public personnel departments are the largest users of tests, and their use has on numerous occasions increased the effectiveness of an employed group. Clerical workers in private industry have also been frequently appraised by standard tests. Large industrial agencies are beginning to use tests for individual guidance in order to determine whether an applicant is well fitted for some position for which he did not apply. Tests have also been used in directing employees toward various courses of training. In the realm of production, standard tests and questionnaires are sometimes used to detect the effect of fatigue, monotony, physical and social working conditions, and payment systems. In the field of merchandising, standardized questionnaires are widely used to ascertain the effects of printed advertisements and radio programs on various groups.

Clinical agencies deal with persons who are mentally abnormal in various degrees. The feeble-minded, the psychotic, the mentally incompatible, the epileptic, the emotional deviate, and the delinquent are studied. A few special tests of moral and neurotic tendencies have been constructed, but the majority of tests applied in the clinical field are designed to aid in determining educational, social, or vocational skills and attitudes.

Civic leaders and agencies are becoming aware of the accuracy of appraisals of public opinion on controversial issues. Rough straw votes of unrepresentative samples of a group of voters have been found to be unreliable, but when a few unambiguous questions of fact or preference are presented, the results are significant. The development of more agencies for appraising attitudes toward economic, social, and political issues is going on rapidly. These will have a marked effect on political actions in a democratic state.

It is interesting to note a marked tendency among the different types of agencies to use the same general types of tests. The school is interested in vocational success, mental health, and the development of character and self-government; hence it uses all types of tests. Industrial agencies find that success in school is one of the best indicators of success in business; so they use many tests of educational achievement as well as tests of character and vocational fitness. Clinical agencies are interested in restoring a person to normal life in school or the community; so they try to evaluate all aspects of a person. All of these agencies are becoming convinced that the whole person has to be considered in any adequate plan for his social development or continued employment.

No one knows exactly how many mental tests are given every year in the United States. The number would certainly amount to several millions. The economic significance is very great in the selection of students for various educational programs and in the selection of employees for various types of work. At present, there are few published reports of the actual savings in dollars and cents.

IV. THE REQUIREMENTS OF A GOOD EXAMINER

The preceding description of the most frequent uses of tests leads to the question, "What are the necessary qualifications of one who is competent to administer and interpret test results?"

There are five main qualifications. First, a good examiner must know why and how to build up a clear set of concepts. Securing this knowledge is often the most difficult part of the training, for concepts, even the very simplest, are hard to understand and to keep from becoming ambiguous. This difficulty is particularly troublesome in the measurement of behavior, for the basic concepts are intangible and still somewhat controversial. The analysis of the processes involved in a response is one of the major persistent problems of psychologists.

Second, a good examiner must be familiar with the best testing instruments. Even to measure the power of a gasoline engine, one must have a well-standardized device for measuring work done in a given period. Since a person responds in many more ways than does an engine, mental measurement is considerably more difficult than measuring a gasoline engine.

Third, a good examiner must know when he has obtained a good sample of performance. Thus, in measuring the power of an engine, the tester must control his experiment. The engine must be run at a standard rate, using standard temperatures and standard fuels and lubricants. Only with such controls can its power be compared accurately with that of other engines using the same standards. In mental measurement, controls are just as necessary and far more difficult to secure. If one wishes to have a good representative sample of a person's speed of reading, it is important not only that he have a good reading test, but also that he control motivation, fatigue, and accuracy of performance.

Fourth, a good examiner must have the ability to judge and use available norms. Norms are figures which show the scores of representative groups. Unfortunately, most of the available norms are from small or specially selected groups. A person who is well trained in measurement will know the special derivation of each set of published results and be able to make allowances for it.

Finally, a good examiner must have ability to report and interpret test findings correctly. If the measurement is part of an experimental procedure, then the findings must be checked with the hypothesis. If the measurement is for individual purposes, then the limits of prediction should be known and specified. Interpretation is one of the most important and difficult parts of the work. A large number of persons who can give tests under standard conditions cannot make accurate interpretations. These persons are useful as laboratory technicians, but they should not be confused with well-trained specialists. The usual training for an expert in the field of mental measurement includes undergraduate emphasis in college upon mathematics, biology, sociology, and at least three years of graduate work specializing in statistical procedures; experimental, systematic, and abnormal psychology; and the theory and application of numerical appraisals of behavior.

V. LIMITATIONS OF TESTING TECHNIQUES

Thus far, the nature of mental measures and some of their uses have been described. It behooves us to become aware of some of their limitations that are often overlooked. Three sorts of limitations are important.

First, mental measurement techniques cannot be expected to make decisions for a person. They can only present the evidence more clearly. From careful measurement, it may appear that a student's chances of average success in a course in electrical engineering, or in second-year French, or in medicine, are only 1 in 100. But the tests cannot decide for a person whether or not he shall attempt the course or the profession. In many cases a person must experience failure in order to be made to realize his limitations.

Second, the best tests available cannot, at present, predict with very great accuracy what a person will do in complex learning or vocational situations. This limitation exists because people vary in motives, emotional balance, social acceptance, health, opportunities, and in many other ways which are not checked in a series of tests. Even from a very complete study of a person, predictions of success ten years later have seldom been very accurate. Most of the predictions of success in college or industry from a particular test or a battery of tests are much better than chance, to be sure, but still they are of little value for individual counseling. Chapter 4 discusses how accurate various predictions of behavior should be in order to be of value to individuals.

Lastly, mental tests ordinarily cannot show why a person made a particular score, but only that he did make the score. This is true of all measures of behavior. For instance, within certain limits, the speed of an automobile can be measured in miles-per-hour, but the speed record does not indicate why it goes so slowly or so rapidly. A very large number of factors, such as engine and chassis design, road and wind friction, fuels, and oils, affect the total speed. The contributory factors in measures of human behavior are considerably more complex. Thus, a speed-of-reading test score depends upon many variables about which one is often uninformed, such as visual acuity, muscular tonicity, verbal information, verbal skills, and desire to succeed. As more careful controls are exercised in testing, the unknown variables will become less important, and the scores more precise.

2

VARIETIES OF APPRAISALS

In the preceding chapter one step in the measurement of behavior was defined as the process of assigning quantitative values to judgments of amounts. This chapter enlarges upon this process by presenting a rapid survey of the varieties of behavior upon which appraisals are commonly made and the ways in which numerical values are combined to provide scores. Such appraisals may be usefully classified according to the degree of control which is exercised over the important variables in a situation. Although the best controlled situations give the most accurate results, useful observations may often be made when the observer is unable to control all of the variables in a situation. Three levels of control may be distinguished: well-controlled, partly controlled, and slightly controlled situations.

I. WELL-CONTROLLED SITUATIONS

Although well-controlled situations are difficult to set up, they may be arranged in a laboratory. There the experimenter can regulate nearly all of the factors which may influence the results. Then a number of factors can be varied according to some definite plan so that the results of the variations can be measured. The best examples of such measurements are found in physics and chemistry laboratories where atmosphere, temperature, pressure, and other variables can be controlled with great accuracy. In biological studies, small animals can be handled with a high degree of control in a laboratory, but there are many variables in human measurement work which are extremely difficult either to control or to evaluate. Three types of tests of behavior stand out as being controlled better than the rest: physiological, psychophysical, and single variable tests.

A. Physiological Measures

Measures of physiological reactions, during which the subject is kept in a very carefully regulated environment, include:

1. Vascular changes – pulse, blood pressure, and volume
2. Breathing – rate, amount, metabolic rate
3. Temperature
4. Electrical disturbances – small changes in potential accompanying muscular and nervous activity
5. Chemical changes – glandular excretions, body fluids, and solids
6. Tensions – mechanical pressures on various organs

The units of measurement in these tests generally show the amount of the force, volume, or direction of a particular change during a particular period of time. These measures are made by persons with a medical or biological training who record the readings of instruments. The instruments have been calibrated by an elaborate series of previous judgments to yield readings in units which are considered to be equal. Such units are described in manuals of physics, chemistry, and engineering. They are not thoroughly discussed in this book, but Chapters 12 and 18 contain evaluations of relationships between a few physiological processes and other sorts of behavior.

B. Psychophysical Measures

Another group of measures which allow fairly rigid controls is concerned with *sensory discrimination.* Measures of sensory discrimination show the results of arousing specific sense organs with carefully prepared stimuli. These may be measures of vision, hearing, touch, pressure, taste, smell, pain, or a large number of sensations of movement, called kinesthetic sensation. The unit of measurement in these studies usually depends upon a judgment of relative speed, size, intensity, or quality. The smallest discrimination that can be regularly made is usually indicated by the number of correct discriminations among a series of standard figures. Samples of psychophysical measures are scores for visual acuity secured from graded charts, and Seashore's (1919) pitch discrimination scores (see Chapter 14).

C. Single Variable Tests

Another possibility of well-controlled measurement is seen when two tests are given, one of which varies from the other by a small difference. This is illustrated by tests of aiming at two sizes of circles (see Figure 1). The test requires a person to place a dot in as many circles as possible during one-minute trials. On this sheet the main difference between the hard and the easy levels of aiming is the size of the circle. Nearly all the other factors, such as distance between the centers and thickness of line, are constant. Persons who have great precision can put a dot in about as many small circles as large circles within a given time. Persons who have very poor precision make much better scores on the large circles than they do on the small circles. Hence, it is possible that a good measure of precision in aiming may be secured by dividing a person's score on the small circles by his score on the larger circles. This division gives a ratio

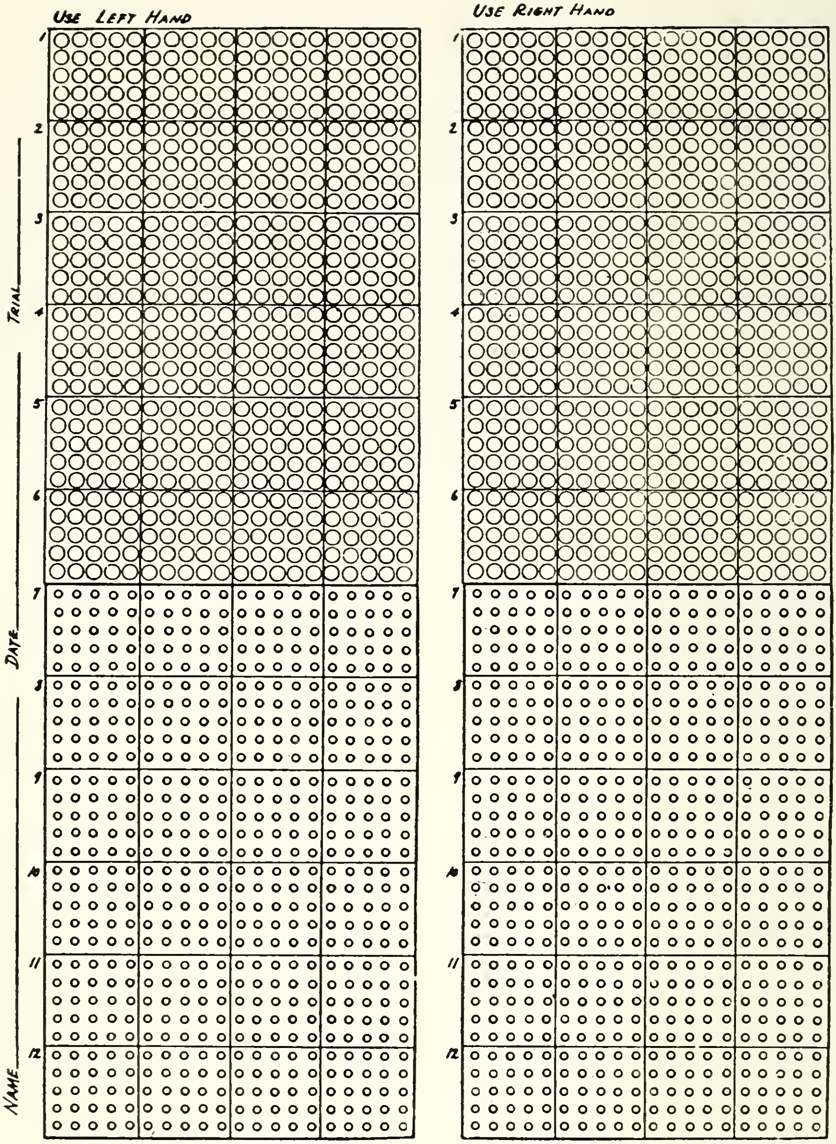

USE LEFT HAND
USE RIGHT HAND
1
2
3
4
5
6
7
8
9
10
11
12
TRIAL
DATE
NAME

FIGURE 1. **Aiming Test, Michigan Nonverbal Series**

(Greene, 1931)

which, if all other factors are constant, depends upon precision of aiming. Ratios of this sort are being studied carefully by a number of investigators in many test situations, such as the comparison of numbers (see Table 1). Some ratios seem to be unusually free from variations caused by different amounts of practice or motivation. A large amount of careful research is still needed to interpret these ratios because, as is well known, they are affected by a person's distribution of energy and temperament, as well as by actual ability for successful work. Not many single-variable tests have been developed as yet, but the technique is widely applicable and one of the most promising.

TABLE 1

Michigan Number Comparison Test

(Form 1)

Name________________________M F Date________Occupation______________

Directions: In some of the pairs below the numbers are the same and in some they are different. If they are the same, do not mark them. If a number in the right hand group is different from the corresponding member of the left hand group, cross it out. Look at Sample A. The numbers of both groups are the same, so do not cross out any. Look at Sample B. The 6 on the right has been crossed out because it is different from the 4 on the left. Do the rest of the samples then wait for the signal "GO" before beginning the main test.

Samples

A	158-158	332-352	798-798	148-048	157-157	952-958	874-854	702-702
B	349-3~~6~~9	461-463	179-179	082-088	054-354	180-180	678-678	567-167
	974-974	587-597	612-612	073-053	251-251	095-055	897-697	153-153
	530-530	640-690	781-781	994-979	925-935	359-359	670-670	STOP

658-668	841-841	530-533	317-318	547-647	975-975	361-461	461-461
913-813	195-195	785-785	702-602	687-087	045-095	361-361	517-517
304-304	832-832	669-069	952-982	587-587	675-675	362-361	028-028
298-248	319-319	512-512	701-711	470-490	867-868	092-092	017-017
041-141	487-487	942-943	870-879	541-551	704-704	991-981	558-558

(*and 30 more rows*)

* * * * * *

Harder Level

254789-254789	894672-894672	123805-123805	349127-349127	692145-692145
390512-390512	734506-934506	364092-364095	678940-671940	470923-470923
317690-357690	475103-475103	697325-697325	581034-581034	784506-784106
673845-672845	956728-956028	623495-623485	017839-017839	814367-514367
380162-380192	178940-178940	147690-137690	578906-578902	678350-678310

534789-534789	978123-968123	123495-127495	456138-456138	312567-412567
845617-843617	345612-345612	567839-567839	789051-789052	672450-672450
258701-298701	890162-890762	127905-127905	456728-456728	345027-345087
012384-012384	016894-016894	512384-512386	819547-819547	031287-031247
567249-567249	708436-708136	467895-407895	946783-946783	895617-395617

(*and 30 more rows*)

(Greene, 1939)

II. PARTLY CONTROLLED SITUATIONS

Loss of control in a test situation comes from inability to regulate all the contributing factors so that a particular score will always indicate a particular pattern of behavior. Partly controlled situations are typical of nearly all measurements of skill, information, and preference. These include both individual and group testing situations.

A. Individual and Group Testing

Individual testing situations are usually better controlled than group situations. They are more useful than group testing situations in the following instances:

1. When oral responses are needed, such as in reading aloud and answering questions
2. In cases where subjects do not readily follow directions or co-operate (This would generally be true of persons below eight years of age, of defective, and of maladjusted persons.)
3. In manipulation of forms or instruments, such as form boards, puzzles, and machine assembly
4. In any situation where it is desirable to evaluate the subject's adjustment to the test situation, such as his persistence, cautiousness, emotional episodes of anger or fear, teasing, balking, or nervous habits
5. In any case where it is desirable and possible to evaluate the subject's methods of work, such as the use of random movements, methodical comparisons, and reasoning

Group tests are more useful than individual tests:

1. When it is desirable to avoid close personal relations with an examiner
2. When it is desirable to compare the effects of group stimulation with the effects of isolation
3. When the subjects can and will co-operate well

4. When the results of group tests have proved to be accurate enough to be useful in a particular situation (Many group tests of written work seem to be as effective as individual tests above the 8-year level.)
5. When economy of time and effort is important

Both individual and group tests are available for many types of behavior. Three general classifications of measurement are found: measures of skill, of retention, and of learning.

B. Measures of Skill

Measures of skill are described under the headings of motor co-ordination, perception, and reasoning.

1. MOTOR CO-ORDINATION. Measures of motor co-ordination seek to evaluate the speed, accuracy, endurance, and

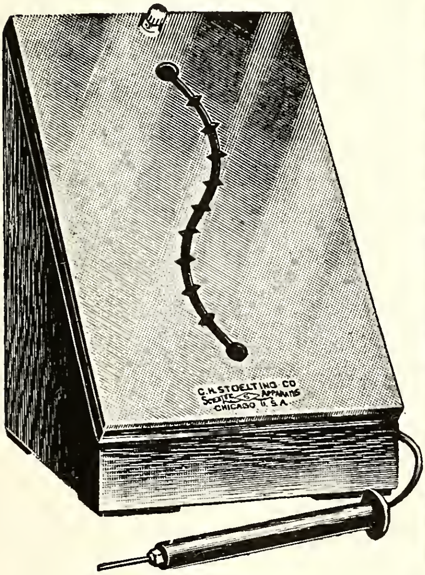


FIGURE 2

Gurnee Tracing Apparatus

(Courtesy of the C. H. Stoelting Co., Chicago, Ill.)

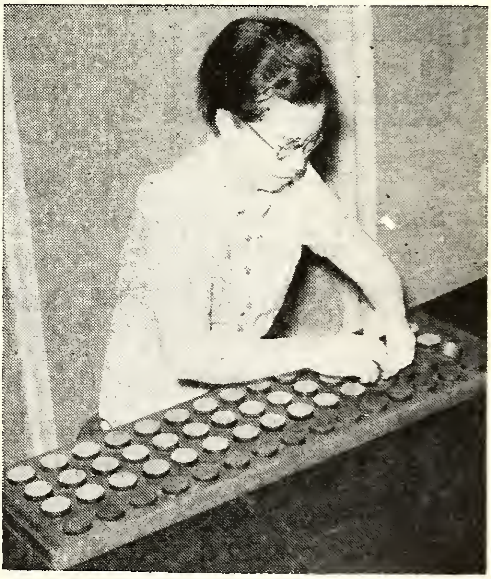


FIGURE 3

The Minnesota Rate of Manipulating Test

Posed by Dr. Bing Chung Ling

(Manufactured by the Educational Test Bureau, Minneapolis, Minn.)

force of movements of particular muscle groups. The units of measurement in this field are usually in terms of amount of work done, such as hits or errors on a target, as in an aiming test, Figure 1; number of contacts in a given time, as in a steadiness test, Figure 2; number of movements completed in a given time, as in various manipulation tests, Figure 3; or the maximum force exerted, Figure 4. Sometimes the amount of work is held constant for all persons, and the score is the time needed to finish the work. Further descriptions of tests of this sort are given in Chapter 12.

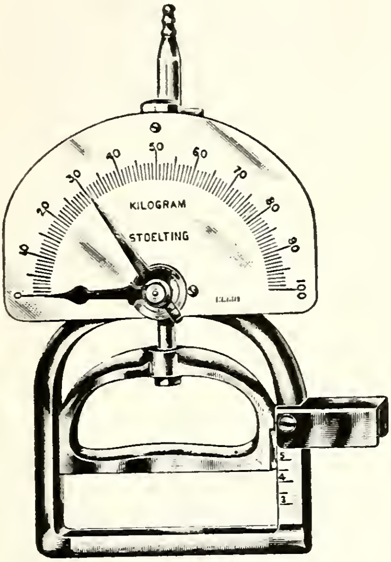


FIGURE 4

Smedley Hand Dynamometer

(Courtesy of the C. H. Stoelting Co., Chicago, Ill.)

2. Perception. Perception tests are indicators of one's discrimination from direct comparison of stimuli, such as sound, chemical, temporal, and spatial patterns. Tests of perception are illustrated by a number and name comparison test, Table 2, and by a form comparison test, Table 64. The unit of measurement in these tests is usually a judgment of same or different, and the raw score is the number of correct judgments in a standard series, usually from three to ten minutes.

Another group of tests which are usually classed as perceptual are called attention span tests. In these, visual stimuli are exposed for a brief period, say one tenth of a second, and then the subject is asked to reproduce or name them. An illustration of attention span tests is McDougall's (1908) Spot-Pattern Test, Figure 5. These tests are not widely used because there is difficulty in controlling lighting and timing. They can best be given in a laboratory.

TABLE 2

Minnesota Vocational Test for Clerical Workers

(Arranged by Dorothy M. Andrew under the direction of
Donald G. Paterson and Howard P. Longstaff)

Name______________________________ **Date**________________________

TEST 1 — Number Comparison

Number Right ________________

Number Wrong ________________

Score = R − W ________________

Percentile Rating ________________

TEST 2 — Name Comparison

Number Right ________________

Number Wrong ________________

Score = R − W ________________

Percentile Rating ________________

Instructions

On the inside pages there are two tests. One of the tests consists of pairs of names and the other of pairs of numbers. If the two names or the two numbers of a pair are *exactly the same* make a check mark (√) on the line between them; if they are *different*, make no mark on that line. When the examiner says "Stop!" draw a line under the last pair at which you have looked.

Samples done correctly of pairs of *Numbers*

79542		79524
5794367	√	5794367

Samples done correctly of pairs of *Names*

John C. Linder		John C. Lender
Investors Syndicate	√	Investors Syndicate

Now try the samples below.

66273894		66273984
527384578		527384578
New York World		New York World
Cargill Grain Co.		Cargil Grain Co.

This is a test for Speed and Accuracy. Work as fast as you can without making mistakes.

Do not turn this page until you are told to begin.

(Copyright, 1933. Reproduced by permission of The Psychological Corporation, 522 Fifth Avenue, New York City.)

More widely used tests, called immediate memory span tests, are sometimes placed in this perception group. In these tests, stimuli are presented one at a time, and the subject is asked to reproduce them exactly as given. Auditory spans for digits, disconnected words, sentences, and bead chains are used in the Stanford-Binet Test, Terman and Merrill (1937). Baker and Leland (1935) developed tests of visual spans for letters and small pictures. Seashore (1919) standardized a test for immediate memory of short tonal phrases. These tests are not dependent solely upon perception of pattern. They also involve a short retention span. It seems likely, however, that good retention over short periods depends upon clear perception of the elements and their relationships. The score in each of these span tests is usually the number of elements recorded correctly during a standard testing situation. Perception tests are described more fully in Chapters 8, 9, 11, 13, and 14. Sometimes the longest series of items which can be reproduced is taken as the score.

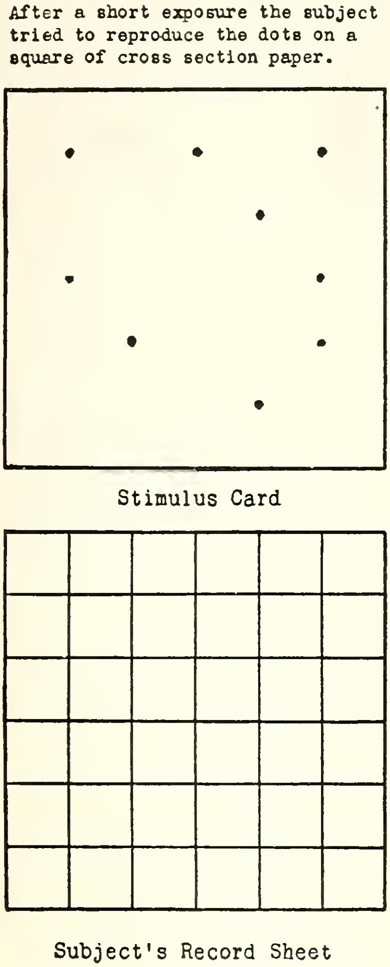

After a short exposure the subject tried to reproduce the dots on a square of cross section paper.
Stimulus Card
Subject's Record Sheet

FIGURE 5

Spot-Pattern Test

After a short exposure the subject tried to reproduce the dots on a square of cross section paper.
(McDougall, 1908)

3. Reasoning. The outstanding characteristic of reasoning is the solution of a problem by the production of a new pattern of behavior out of experiences which were previously

arranged in other patterns. Under reasoning are included tests of inference and of problem solving.

In inference tests the subject is asked either to detect inconsistencies or to make inferences of his own. In the simplest form, two objects or statements are presented and the subject must show their relationships by comparing each with a third. This form is seen in a syllogism test, Table 3, in which the subject is asked to check those inferences which must follow from preliminary statements. Another illustration is a speed-of-reading test, page 224, in which the subject is to find a word in the last part of a paragraph which spoils the meaning of the first part. A more complex form of reasoning test is shown in the interpretation of scientific data, Table 4c. There the subject is asked to check the inferences which may be correctly drawn from an account of a situation where there are four or five variables involved. The unit of measurement in these tests is usually a judgment of *true, false,* or *undetermined* relationship.

TABLE 3

Syllogism Test

Directions: In each section below, read the first sentence and the line marked a. If what a says follows from the first sentence, make a circle around the T in line a. If what a says is false according to the first sentence, circle the F. If the statement in a need not follow from the first sentence, circle the Q. Do the same for lines b and c. The sample below is marked correctly.

Sample: All good dancers dance frequently; the men in this house do not dance frequently; therefore, —

a. The men in this house are not good dancers [✓ T] F Q
b. All frequent dancers are good dancers T F [✓ Q]
c. Some good dancers are in this house T [✓ F] Q

Form A.

1. All the people living on this farm are related to the Joneses; these old men live on this farm; therefore, —

a. These old men are related to the Joneses. T F Q
b. All the people related to the Joneses are these old men. T F Q
c. Some people related to the Joneses are not these old men. T F Q

(After Wilkins, 1928)

TABLE 4

Cooperative Chemistry Test

A. SELECTION OF FACTS

Directions: Following are a number of incomplete statements each of which may be completed by one or more of the words or phrases given below the statement. Place a plus sign (+) in the parentheses after those words or phrases which will make the statement true, as in the following sample.

Sample: Oxygen is an element which:

a. Acts chemically as a metal . . . . ( ) a.
b. Unites with hydrogen forming water . . . . (+) b.
c. Is a good conductor of electricity . . . . ( ) c.
d. Is rarely found in nature . . . . ( ) d.
e. Supports combustion . . . . (+) e.

B. TERMINOLOGY

Directions: Below is a numbered list of chemical terms arranged in alphabetical order. Following the list are several definitions or descriptions of terms used in Chemistry. Read each definition or description, decide what term it is, then place the number of the term in the parentheses after the definition or description.

1. Acid salt
2. Aliphatic compounds
3. Amalgam
4. Anhydrous
5. Aromatic compounds
6. Atom
7. Atomic number
8. Basic salt
9. Carboxyl
10. Critical temperature
11. Esters
12. Heat of solution
13. Hydrogenation
14. Inversion
15. Invertase
16. Ketone
17. Kindling temperature
18. Metalloids
19. Molar solution
20. Mole
21. Molecule
22. Monel metal
23. Normal salt
24. Osmosis
25. Saturated solution
26. Solute
27. Solvent
28. Spontaneous combustion
29. Standard solution
30. Zymase

a. Elements which possess in some degree the physical properties of metals and the chemical properties of nonmetals . . . . ( ) a.
b. Characterized by one or more six carbon atom rings . . . . . . ( ) b.
c. A compound composed of a negative ion other than hydroxyl or oxygen in combination with some positive ion other than hydrogen . . . . ( ) c.
d. Characterized by an open chain of carbon atoms . . . . ( ) d.
e. A mixture of mercury and one or more other metals . . . . ( ) e.
f. Process by which sucrose is changed into a mixture of equal parts of glucose and fructose . . . . ( ) f.
g. Characteristic of substances which are not combined with water . . ( ) g.
h. A group of elements which characterize the organic acids . . . . ( ) h.
i. Burning produced by slow oxidation and the accumulation of heat . ( ) i.
j. Smallest unit of a substance which takes part in ordinary chemical changes . . . . ( ) j.
k. The temperature at which a substance begins to glow or bursts into a flame . . . . ( ) k.

TABLE 4 — *Continued*

Cooperative Chemistry Test

C. APPLICATION OF PRINCIPLES

Directions: In each of the following exercises a problem is given. Below each problem are two lists of statements. The first list contains statements which can be used to answer the problem. Place a plus sign (+) in the parentheses after the statement or statements which answer the problem. The second list contains statements which can be used to explain the right answers. Place a plus sign (+) in the parentheses after the statement or statements which give the reasons for the right answers. Some of the other statements are true but do not explain the right answers; do not check these. In doing these exercises then, you are to place a plus sign (+) in the parentheses after the statements which answer the problem and which give the reasons for the RIGHT answers.

Sample: Coal gas which has not been previously mixed with air is burned at a gas jet. At another similar gas jet the coal gas is mixed with air before it is burned. Will there be any difference in the amount of light given off by the flames of the two gas jets? Why? If a cool aluminum pan is placed over each flame will there be any difference in the amount of soot deposited on the pan in the two cases? Why?

The flame at the first gas jet will give off:

a. More light than the flame at the second gas jet . . . . (+) a.
b. The same amount of light as the flame at the second gas jet . . . ( ) b.
c. Less light than the flame at the second jet . . . . ( ) c.

The soot deposited by the first gas jet will be:

d. More than that deposited by the second gas jet . . . . (+) d.
e. Less than that deposited by the second gas jet . . . . ( ) e.

Check the following statements which give the reason for the answer or answers you checked above.

f. Incomplete combustion leaves some uncombined carbon in the flame . (+) f.
g. The presence of nitrogen retards combustion . . . . ( ) g.
h. Particles of uncombined carbon glow when heated . . . . (+) h.
i. Combustion is more complete in the first flame . . . . ( ) i.
j. The amount of air mixed with the gas does not affect the amount of light produced by the burning gas . . . . ( ) j.
k. Some uncombined carbon in a flame is deposited on a cool surface placed in a flame . . . . (+) k.

In the above exercise, the statements which answer the problem are a and d. These statements are checked because they tell what would be likely to happen. Statements f, h, and k have also been checked because they are reasons which help to explain why a and d would happen. You will notice that statement g is a true statement but it does not give a reason for any of the right answers; hence it is not checked. The other statements are not true and are not checked. Do each exercise in this way.

(Samples from Hendricks *et al.*, 1934. By permission of the Cooperative Test Service, Inc.)

Other fairly common forms of tests which involve reasoning are analogies, page 305, and disarranged sentences, page 303. In both of these, isolated words or figures are to be matched or arranged according to relationships indicated in a context.

In problem-solving tests a person must become aware of a problem, select a hypothesis or plan for solution, and then apply this plan to see whether it will succeed. If it succeeds, he scores a point; if it does not, he must discard it and seek

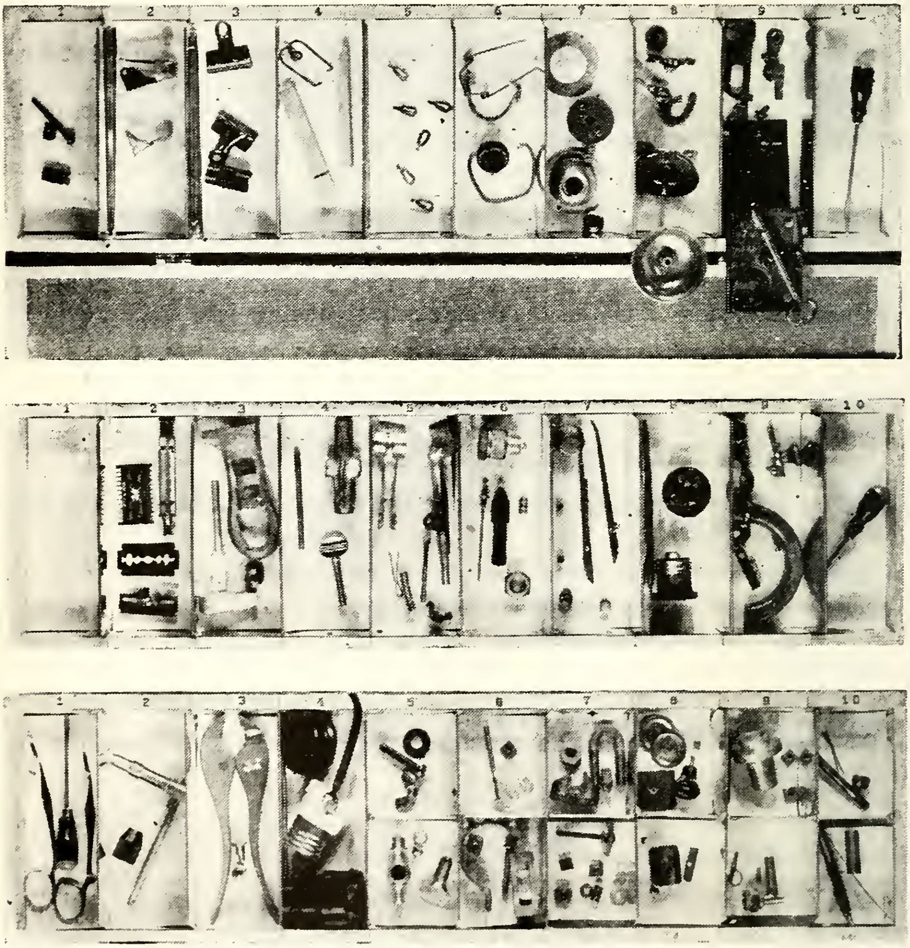


FIGURE 6. Minnesota Mechanical Assembly Test Materials

Top, Box A; Center, Box B; Bottom, Box C

(Paterson *et al.*, 1930. By permission of the Marietta Apparatus Co.)

another solution. Illustrations of this type of test are seen in (*a*) mathematical problems, page 300; (*b*) the assembly of apparatus, Figure 6; (*c*) pencil mazes, Figure 52, page 353, in which a person can see the whole maze at once and evolve a plan for finding a way through it; and (*d*) many verbal situations in which one must try out several hypotheses. The unit of measurement is a problem or part of a problem which is solved. Sometimes the method of solution and the time needed are also noted.

Reasoning is common in ordinary life, but reasoning ability is very hard to measure on a quantitative basis, for different persons have different amounts of information which aid in the solution of the problem. Some persons may know the solution before they are asked the problem. It is necessary to eliminate differences in relevant information in order to evaluate reasoning activities. This elimination may be accomplished with fair success in verbal tests by limiting the vocabulary to words which are very familiar to all in the group being tested.

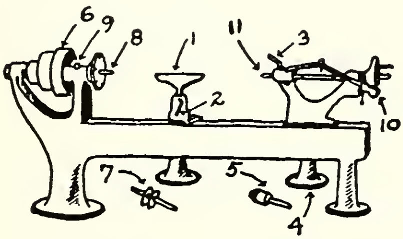
C. Measures of Retention

In tests of retention a person is asked to show how many and what sorts of items he can recall or recognize when a stimulus is given. Sometimes the person's feelings or emotions affect the recall process, so that two kinds of tests have been devised: (*a*) information tests, which are designed to be free from emotional factors, and (*b*) free association tests, which are intended to disclose emotional experiences or moods as well as information.

1. INFORMATION TESTS. Information tests require recall or recognition of verbal or other material. In many ways these tests are the best developed and the most important in use today. The best controlled are probably the recognition tests, in which the examinee chooses the most appropriate answer from two or more which are presented. This type is illustrated in Table 4A, in which statements are to be

marked true or not marked at all, and by Table 4B, in which one of thirty answers is to be chosen. All of these items could also be cast into a recall type, in which the examinee must supply the answer. Table 18, page 110, shows a completion test in which a particular word must be recalled to complete a sentence. On page 302 the opposite of the stimulus word is to be given.

Information tests are not limited to words; there are many good pictorial forms. Picture-naming tests are widely used for appraisal of preschool and mechanical vocabularies. Picture tests are often superior to word tests when knowledge of the use or motion of parts of a machine is to be tested. Figure 7 shows a picture of a lathe, the parts of which are numbered. The subject is asked to indicate which parts perform certain functions. In nearly all information tests the unit of measurement is an item correctly answered and the raw score is the total number of correct items. Sometimes corrections for chance successes are applied as described in Chapter 6. Information tests form the chief component of many of the tests which are described in Part II.



22 CARRIES BELT---------- (   )
23 ADJUSTS HEIGHT OF REST--- (   )
24 SUPPORTS TOOL-------- (   )
25 OILS BEARING--------- (   )
26 FASTENS TAIL STOCK CENTER- (   )
27 REVOLVES WORK-------- (   )
28 ADJUSTS TAIL STOCK CENTER- (   )
29 HOLDS WORK---------- (   )
30 FASTENS TAIL STOCK----- (   )
31 HOLDS DRILL---------- (   )
32 HOLDS EMERY WHEEL---- (   )

FIGURE 7. **Lathe Items from the Detroit Mechanical Aptitudes Test for Boys**

(Baker and Crockett, 1928. By permission of the authors and the Public School Publishing Co.)

2. Free Association Tests. In these tests the information called for is not carefully limited to certain answers, but it is purposely left vague. Thus, in the Kent-Rosanoff Free Association Test, Table 5, a person is asked to respond to a

TABLE 5

The Kent-Rosanoff Free Association Words

1. Table
2. Dark
3. Music
4. Sickness
5. Man
6. Deep
7. Soft
8. Eating
9. Mountain
10. House
11. Black
12. Mutton
13. Comfort
14. Hand
15. Short
16. Fruit
17. Butterfly
18. Smooth
19. Command
20. Chair
21. Sweet
22. Whistle
23. Woman
24. Cold
25. Slow
26. Wish
27. River
28. White
29. Beautiful
30. Window
31. Rough
32. Citizen
33. Foot
34. Spider
35. Needle
36. Red
37. Sleep
38. Anger
39. Carpet
40. Girl
41. High
42. Working
43. Sour
44. Earth
45. Trouble
46. Soldier
47. Cabbage
48. Hard
49. Eagle
50. Stomach
51. Stem
52. Lamp
53. Dream
54. Yellow
55. Bread
56. Justice
57. Boy
58. Light
59. Health
60. Bible
61. Memory
62. Sheep
63. Bath
64. Cottage
65. Swift
66. Blue
67. Hungry
68. Priest
69. Ocean
70. Head
71. Stove
72. Long
73. Religion
74. Whiskey
75. Child
76. Bitter
77. Hammer
78. Thirsty
79. City
80. Square
81. Butter
82. Doctor
83. Loud
84. Thief
85. Lion
86. Joy
87. Bed
88. Heavy
89. Tobacco
90. Baby
91 Moon
92. Scissors
93. Quiet
94. Green
95. Salt
96. Street
97. King
98. Cheese
99. Blossom
100. Afraid

(Rosanoff, 1920. By permission of John Wiley and Sons, Inc., New York.)

stimulus word with the first word that occurs to him. His answer depends not only on what words he knows, but also on his mood and his emotional reactions to situations represented by the words. Another test of this sort is the Rorschach Test (1921), Figure 8, in which one is asked to tell all the ideas suggested by ten ink blots in turn. The units of measurement are the number of answers which conform to a standard previously determined by tabulating results from various groups. These standards are at present qualitative descriptions rather than quantitative scales, but quantitative standardizations are developing. Chapter 17 describes measures of this sort more fully.

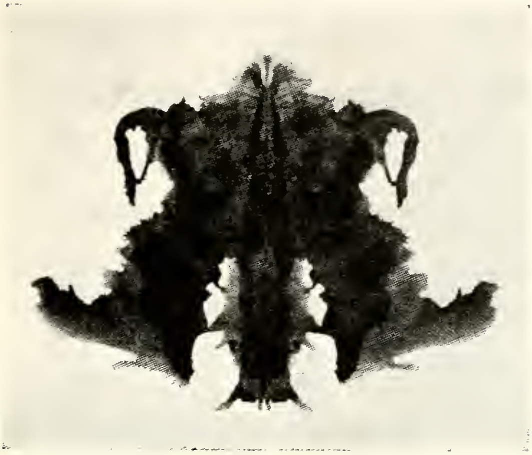


FIGURE 8

A Rorschach Test Figure

About One-Sixth Actual Size

(By permission of Verlag Hans Huber, Berne, Switzerland

D. Measures of Learning

Although learning often takes place during a test, and many definitions of intelligence include learning as its most important element, tests designed to measure learning are not common. This lack of tests of learning is due to the slowness with which learning usually takes place and the many serious difficulties encountered in measuring it. Measures of practice from repeated tests are described in Chapter 20.

Tests which are doubtless affected by speed of learning include the digit-symbol test of the United States Army Beta, page 314. Here a number is to be written according to a code below each symbol printed on the page. Other examples of this type of test are the indexing and classification tests shown in Table 6, and various tests of code writing.

TABLE 6

O'Rourke Clerical Aptitude Test, Junior Grade, (Clerical Problems)

Form 1

O'Rourke Series of Aids in Placement and Guidance

PRINT YOUR NAME .... (Last name) (First name) (Middle initial)
Date.................. 19...... (Month) (Day) (Year)
Age now.............. (Years) (Months)
Write the last school grade you COMPLETED..........................
Age at that time..................

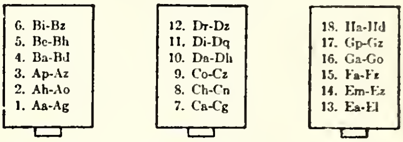
You will have five minutes to study the samples on this page. The tests on the following pages are like these. Be sure you understand the samples.

SAMPLES

Time commenced..........................

FILE DRAWERS

Sample I. You are to file alphabetically in the file drawers shown at the right. Each drawer contains six folders. After each of the names listed below the third file drawer, you are to write the number of the folder in which that name should be filed.



The first name, "Appel," should be filed in the folder for names from Ap to Az. This folder is numbered 3, so 3 is written after the name "Appel". The second name, "Cecil," should be filed between Ca and Cg, which is folder No. 7, so 7 is written after it. "Denby" should be filed in folder "Da-Dh", so WRITE 10 AFTER "DENBY". Next write the number of the folder in which "Earl" should be filed. The number of the correct folder is 13.

Name	**Folder**
1. Appel	3
2. Cecil	7
3. Denby	
4. Earl	

Sample II. Write a C after the name of each man in the list at tha right who is a teacher, is between 25 and 40 years of age, and resides in either Indiana or Illinois.

Name	*Age*	*Occupation*	*Residence*	
Beach, R. P.	38	Teacher	Michigan	
Mark, J. K.	33	Teacher	Indiana	C
Savoy, W. T.	23	Student	Indiana	
Bard, T. I.	31	Teacher	Illinois	

Nothing is written after Mr. Beach's name, as his residence is neither Indiana nor Illinois. C is written after Mr. Mark's name, as he is a teacher, is between 25 and 40, and resides in Indiana. You should not write anything after Mr. Savoy's name, as he is not a teacher. Write C after Mr. Bard's name as he is a teacher, is 31 years old, and is from Illinois.

Sample III. If the name, the address, and the charge are not exactly the same in the copy as in the original, X is to be written on the line at the right. If the copy is the same as the original, write S.

Original			**Copy**			
Name	*Address*	*Charge*	*Name*	*Address*	*Charge*	**Check here**
Harris, E. G.	38 Oak St.,	$38.95	*Harris, E. G.*	*26 Oak St.,*	*$38.95*	X
Graves, L. R.	21 Rex Ave.,	$16.16	*Graves, L. R.*	*21 Rex Ave.,*	*$16.16*	S
Overt, P. W.	15 End St.,	$ 5.00	*Overt, P. W.*	*15 End St.,*	*$10.00*	

The first is marked X because the address was copied incorrectly. S is written after the second, as the name, address, and charge were all copied exactly as in the original. In the third, a mistake was made in copying the charge, so WRITE X ON THE LINE.

DO NOT TURN THIS PAGE

(By permission of the Psychological Institute, Washington, D. C.)

All of these tests require the learning of associations. The unit of measurement is the correct placement of a symbol according to the code. The total score depends upon the speed and accuracy with which symbols are placed. In laboratory situations speed of learning with mirror-drawing, mazes, and nonsense syllables has been extensively studied. In school situations measures of the acquisition of language and mathematical skills have been given much attention. These studies result in estimates of speed of learning, but few standardized tests have appeared.

III. SLIGHTLY CONTROLLED SITUATIONS

One is often interested in securing accurate records of what persons do in usual situations in which the experimenter is not trying to exert control. The presence of the experimenter often inhibits responses which a child might make if the experimenter were not present or not noticed. Studies of honesty, dominance, sociability, and social patterns of response of other sorts are probably more conclusive when uncontrolled situations are observed. This is also true of emotional episodes, such as outbursts of anger, fear, or jealousy. In a study of small children comparisons of new responses of language and other forms of adaptation may often be studied better when the experimenter is not introducing an unusual element into the situation. Slightly controlled methods also have rather wide application in the study of adults in usual situations, such as studies of responses to advertisements and propaganda. In this division are included time-sampling, ratings, logs, interviews, and case histories.

A. Time-Sampling

Observations of uncontrolled situations are generally made by what is known as a time-sampling method. Here observers take samples of performance at regular intervals during a day or over longer periods of time. The results show the frequency with which types of behavior patterns appear in certain situations. A disadvantage of this method is the difficulty of getting observers to agree upon reports of what they have seen. Even with considerable training there appear marked discrepancies in some reports. When observers do agree, there is some evidence that they have eliminated from their reports the controversial data upon which they could not agree. Such reports are not complete and may fail to show true relationships. This same difficulty is present in all kinds of observation, but it seems to be more serious in time-sampling, with the preconceived ideas of the observer serving to direct the attention toward certain types of behavior. In spite of these difficulties, Jersild and Meigs (1939) have summarized studies where the results are very consistent and fairly complete. The unit of measurement is usually the appearance of the noticed type of behavior, recorded as being present at one of the regular intervals of sampling. The raw score is the number and duration of the periods of each activity under consideration (see Chapter 18).

B. Ratings

For the evaluation of typical behavior and of attitudes and interests, a very large number of rating scales have been developed. Some of these use self-ratings and some use ratings by others. Self-ratings are considered to be one of the best ways of evaluating interests and attitudes of adults. Illustrations of ratings are seen in Table 79, page 461, in which vocational preferences are listed; in Table 7, in which attitudes toward the church are to be appraised; and in Figure 9, in which a person expresses a preference among artistic materials. The unit of measurement is based upon

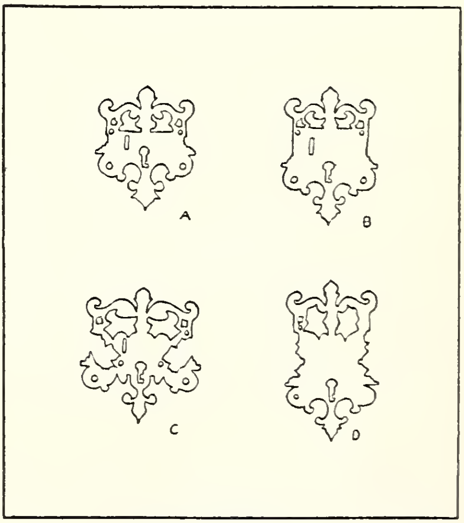


FIGURE 9

Sample of the Artistic Preference Test

(McAdory, 1929, p. 6. By permission of the Bureau of Publications, Teachers College, Columbia University.)

Thurstone

TABLE 7

Attitude toward the Church

Check every statement with which you fully agree:

1. I think the church is a divine institution, and it commands my highest loyalty and respect.
3. I feel the good done by the church is not worth the money and energy spent on it.
5. I believe that the church is losing ground as education advances.
7. The teaching of the church is altogether too superficial to be of interest to me.
9. I think the church has a most important influence in the development of moral habits and attitudes.
11. I regard the church as a harmful institution, breeding narrow mindedness, fanaticism, intolerance.
13. I believe in the ideals of my church, but I am tired of its denominationalism.
15. I'm not much against the church, but when I cannot agree with its leaders I stay away.
17. I believe that the church practices the Golden Rule fairly well and has a consequent good influence.
19. I feel the church is ridiculous, for it cannot give examples of what it preaches.
21. My church is the primary guiding influence in my life.
23. My attitude toward the church is one of neglect due to lack of interest.
25. I am sympathetic toward the church, but I am not active in its work.
27. I know too little about any church to express an opinion.
29. I am slightly prejudiced against the church and attend only on special occasions.
31. There is much wrong in my church, but I feel it is so important that it is my duty to help improve it.
33. I think the church is unreservedly stupid and futile.
35. I feel the church is petty, easily disturbed by matters of little importance.
37. I believe the church is non-scientific, depending for its influence upon fear of God and hell.
39. It seems absurd to me for a thinking man to be interested in the church.
41. I believe that anyone who will work in a modern church will appreciate its indispensable value.
43. My attitude toward the church is passive, with a slight tendency to disfavor it.
45. I have nothing but contempt for the church.

(Abbreviated from Thurstone and Chave, 1929, p. 23. By permission of The University of Chicago Press.)

a judgment of like, indifference, or dislike for a particular activity, or upon a choice between items. Raw scores show total likes or dislikes and also particular regions of interest.

Self-ratings are often considered to be invalid owing to the difficulty that persons have in evaluating their feelings accurately, and to the fact that in some cases they may purposely falsify the reports in order to make the results more in line with what is socially acceptable. Ratings are also affected by ambiguities in definitions, as when persons are asked to evaluate rather complex traits, such as tact or reasoning. Lastly, ratings are in some cases subject to halo effects. A halo effect occurs when one general classification, such as attractive or unattractive, influences the rater's judgment on a large number of other traits of a more specific nature, such as neatness, promptness, kindness, and intelligence. The refinement of ratings is discussed in Chapters 16 and 23.

C. Logs

Another method of recording behavior is a log. This differs from direct observation in that the observer records pertinent outstanding events over a considerable period, such as an hour or a whole day. In many cases, outstanding events are significant of behavior, and for certain purposes, logs are valuable. Teachers' logs and camp counselors' logs are often used to check results of tests and ratings (see Chapter 18).

One disadvantage of this method is the fact that it is dependent upon emotional and subjective variables in the observer. If the observer is feeling particularly well and happy, he will probably not record as many annoying episodes as when he is suffering from a severe headache. Unless the log is recorded very regularly and with considerable attention to the methods of recording, it is extremely difficult to summarize. In many cases logs have proved to be so vague as to be useless for comparing persons or groups.

D. Interviews

Because interviewing represents an exceedingly adaptable method of securing data, it is more useful than some of the other methods. It allows the interviewer to ask questions in such a way as to secure the confidence of the person who is being interviewed. This is often difficult in written or oral tests because in many cases individuals are on guard and unwilling to show exactly what they can do or what they are thinking. With a well-defined outline for interviewing, many aspects of behavior can be systematically covered and a fair appraisal made in a very short time (see Chapter 17).

On the other hand, interviews have, in addition to the function of appraising a person, the function of giving information and suggestions, and it is often quite difficult to avoid suggesting, by leading questions, the answers which one wishes to get. Interviews have another disadvantage, that of being relatively unstandardized. The results are difficult to handle numerically unless a standard rating technique has been used. Interviews are used for nearly all types of employment, counseling, clinical, and psychoanalytical appraisals.

E. Case Histories and Biographies

Case histories and biographies are among the most useful sorts of records for evaluating long-time trends in a person's development. A careful case history will include a record of family history, physical development, health, tests of progress in intellectual pursuits, social and economic adjustments, and emotional patterns of development. It will try to show how much these aspects were dependent upon environment and how they were related to one another. It is on the basis of careful case histories and tests that the most accurate predictions can be made.

One disadvantage of the case-history method is the great difficulty of getting a true record of events which occurred several years previously. Case histories usually include rather vague memories of a person, his friends, parents, and teachers. These memories are subject to forgetting and improvising processes. Case histories of the same person made up by two investigators sometimes differ in very important aspects. The best case histories are those which have been started at an early age and thereafter continued by additions at regular periods.

3

THE INTERPRETATION OF SCORES

The previous chapters have described ways of sampling a person's behavior by recording his responses to particular types of test items. The total number of credited responses, called a *raw score,* represents the performance of a person in a test situation. The raw score, however, is of little use by itself. For instance, no one becomes much wiser by learning that Frank's score in a vocabulary test was 20 words correctly defined. The raw score must be interpreted by comparing it with other scores, and two sorts of comparisons are commonly made. One procedure compares Frank's score with the average scores of various age groups, and thus gives an indication of Frank's relative maturity. This procedure is described in Chapter 10. The other procedure compares Frank's score with scores of other persons to indicate his relative position in a particular group. One of the most important facts to know about a person is his position in some standard group, for this shows how well he is equipped to meet competition. This chapter describes several common ways of changing raw scores into indications of one's relative position in a group. Furthermore, it shows how groups can be measured and compared.

I. THE LIMITS OF A SCORE

Since a score on a test is only an approximate indicator of the results of an observation, it has somewhat variable limits, of which two are commonly defined.

In one case a score is taken to represent an achievement which a person has demonstrated. Thus, the age of 50 years is interpreted to mean that a person has completed 50 years of life, and a group of 50-year-olds is usually thought of as ranging in age from 50.0 years to 50.999+ years. The midpoint of these ages is 50.5 years. Many scores of time, distance, and mental measurement are interpreted in this fashion. This meaning of a score as the lower limit of steps in a scale is in common use. A recent check by the author shows that nearly 30 per cent of reports of appraisals of persons in current journals use this interpretation, but there seems to be an increasing use of the second interpretation because it gives a truer approximation of the facts.

In the other case, a score is considered to be a midpoint of a scale unit. In measuring the height of a person many times with a ruler, one is likely to make errors and to record values above the true length as often as below it. Hence, any obtained score such as 50 is thought of as representing values from 49.5000 to 50.49999. The midpoint of this unit is 50.0, which is .5 below the midpoint secured through the use of the first interpretation. Although either of these two interpretations of a score is usually acceptable, changing from one to the other during a calculation will, of course, introduce errors.

II. FREQUENCY OR DISTRIBUTION TABLES

The immediate result of testing a group of persons is just a pile of corrected test sheets, or a list of scores, as in Table 8. In order to compare these scores, it is convenient to arrange them in a *frequency table.* Such a table shows the number of times each score was made. The number of persons who made a particular score is called the *frequency* of that score. Table 9 illustrates the steps usually followed in making a frequency table. First, one finds the *range of scores,* which is the difference between the highest and lowest scores in a group. If the range is large, it is not advisable to count the number of times each score was made. Unnecessary work is eliminated by tabulating scores in small groups, called *class intervals.* It has been found that 10 class intervals allow one to calculate group norms about as accurately as a larger number of class intervals would permit. Therefore, the range of scores is divided by 10 to find the size of the class interval. For instance, the total range of scores in Table 8 is 44 (71–27), and this range divided by 10 is 4.4. To simplify tabulations the next highest whole number, 5, is taken as the class interval to be used. The scores are then written as in Table 9, Column 1, showing the limits of each class interval. Following a common usage for scales of distances or time, the lower limit of each class interval is often used alone as in Table 9, Column 2. These limits are usually selected so that they are multiples of the interval chosen. The smallest scores are put at the lower end of the column, and the highest at the upper end. The number of persons whose scores are found in a class interval is indicated by the tab marks in Column 3, and by the frequencies in Column 4.

III. HISTOGRAMS AND FREQUENCY CURVES

It is often easier to understand distribution tables if they are pictured, and one of the best ways of picturing them is by a *histogram.* A histogram is drawn on graph paper with the base line divided into class intervals, and a standard area above this line allotted to each person. Thus, in Figure 10 each person is represented by one small rectangle. The whole figure represents the data that are shown in Table 9. A *frequency curve* is a histogram which has been smoothed

TABLE 8

Raw Scores Made by 100 Students on a Number Comparison Test

35	40	50	43	60	53	45	50	71	43
45	60	36	60	55	47	42	45	55	52
46	44	51	62	44	52	33	44	63	37
39	62	66	39	59	58	43	44	38	44
56	39	47	41	40	36	40	48	55	58
50	47	45	42	50	30	39	33	51	36
57	60	32	69	53	50	55	53	48	27
40	47	35	58	63	53	64	35	54	49
45	62	45	33	43	61	64	46	42	49
53	57	54	59	48	30	50	49	29	37

Number of Scores (N) = 100
Sum of Scores (Σ) = 4796
Mean (M) = 47.96

TABLE 9

Frequency Table of Scores in Table 8

Class Interval of Five

(1) *Score*	(2) *Score*	(3) *Tabs*	(4) *Frequencies*	(5) *Cumulative Frequencies*
70–74.9	70	1	1	100
65–69.9	65	11	2	99
60–64.9	60	𝍸 𝍸 11	12	97
55–59.9	55	𝍸 𝍸 11	12	85
50–54.9	50	𝍸 𝍸 𝍸 11	17	73
45–49.9	45	𝍸 𝍸 𝍸 111	18	56
40–44.9	40	𝍸 𝍸 𝍸 11	17	38
35–39.9	35	𝍸 𝍸 111	13	21
30–34.9	30	𝍸 1	6	8
25–29.9	25	11	2	2

N is the sum of column (4) 100

Calculation of quartiles and centiles

P_{75} = 75th centile = 3rd quartile = $55 + (2/12)5 = 55.83$
P_{50} = 50th centile = median = $45 + (12/18)5 = 48.33$
P_{25} = 25th centile = 1st quartile = $40 + (4/17)5 = 41.17$

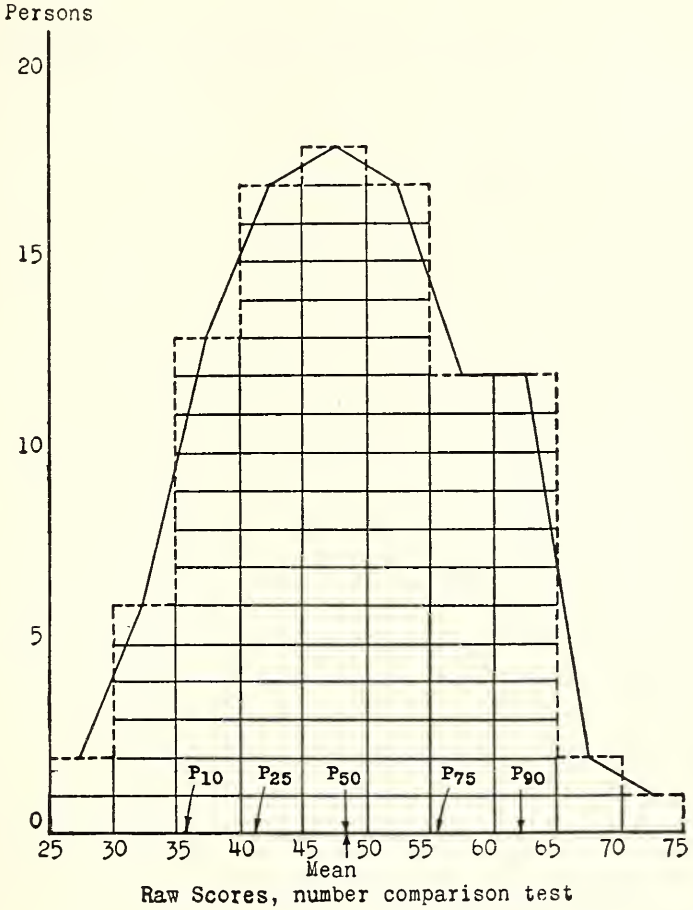
$$Q = \frac{P_{75} - P_{25}}{2} = \frac{55.83 - 41.17}{2} = 7.33$$

P_{90} = 90th centile = $60 + (5/12)5 = 62.08$
P_{10} = 10th centile = $35 + (2/13)5 = 35.77$

$$\text{Sk} = \text{Skewness} = \left(\frac{P_{90} + P_{10}}{2}\right) - P_{50} = \frac{97.85}{2} - 48.33 = .60$$

$$\text{Ku} = \text{Kurtosis} = \frac{Q}{P_{90} - P_{10}} = \frac{7.33}{26.31} = .278$$

according to some method, frequently by connecting the midpoints of adjoining columns, as in Figure 10. A frequency curve is supposed to be a little truer picture of a distribution



Note: All numbers refer to the lower limit of their class intervals.

FIGURE 10. **Histogram and Frequency Curve of Scores in Table 8**

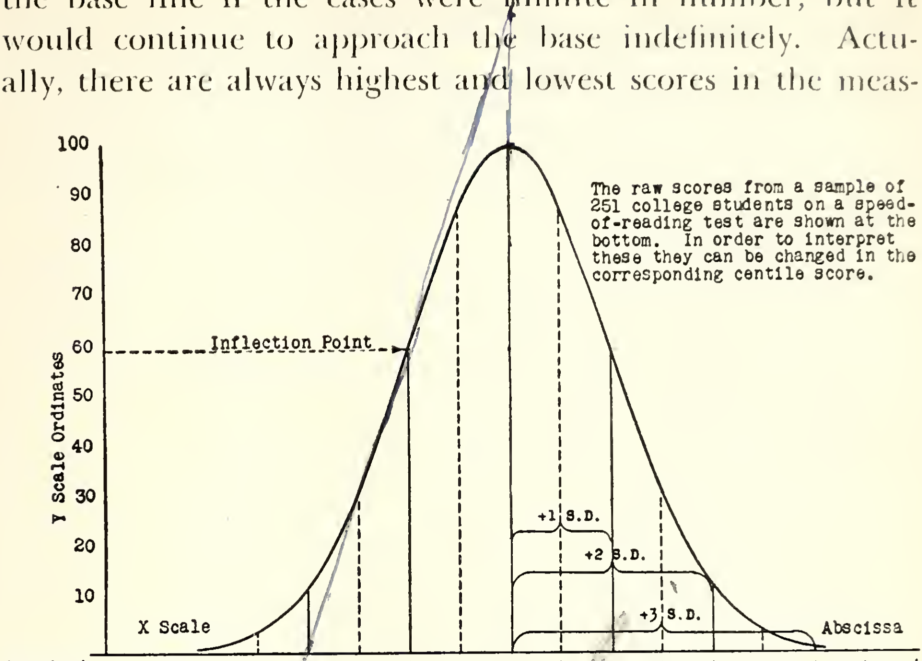
than a histogram, since the cases in a class interval are seldom actually distributed evenly over the whole interval, as they are represented in a histogram. On close inspection, they are found to be distributed unevenly, with more cases falling at that end of the interval which is nearer the center of the whole group. When a frequency curve has a large number of class intervals, and represents a large number of cases, the surface will be very smooth indeed. It can then be considered to be a continuous curved line, as in Figure 11.

IV. NORMAL FREQUENCY CURVE

Frequency curves have been compiled for literally thousands of cases in hundreds of biological, physical, and psychological studies, and a very striking similarity among these curves is the rule. A large number of them closely resemble what is known as the *normal curve of distribution* or the *Gaussian curve*. The normal curve, Figure 11, has been found to be typical in many instances of large samples of:

1. Mechanical variations, such as the distribution of heads and tails shown after tossing coins, combinations of playing cards, and numbers in wheels of chance
2. Mathematical combinations, such as polynomial expansions
3. Physiological structures: height, weight, pulse rate, and ossification
4. Social phenomena: number of books in private homes and attitudes toward various social and political institutions
5. Mental abilities: information and skills in school subjects, intelligence tests, occupational skills, and errors in observation

The shape of this curve, which has been studied minutely, is worth noting, for its characteristics have led to the development of an important scaling technique. First, observe that the greatest frequency of scores is at the midpoint. From this midpoint the sides slope down, slowly at first, then more rapidly until a point in the curve, called the *inflection point*, is reached, whence the curve flattens out gradually. The curve continues until it becomes almost parallel with the base line. Theoretically, the curve would never touch the base line if the cases were infinite in number, but it would continue to approach the base indefinitely. Actually, there are always highest and lowest scores in the measurement of individuals in a real group. The normal curve, whose algebraic dimensions are constant and well known, has become a standard with which group measures may be compared.



The raw scores from a sample of 251 college students on a speed-of-reading test are shown at the bottom. In order to interpret these they can be changed in the corresponding centile score.

Standard deviation		-3.5	-3.0	-2.5	-2.0	-1.5	-1.0	-0.5	0.	+0.5	+1.0	+1.5	+2.0	+2.5	+3.0	+3.5
Standard scores		15	20	25	30	35	40	45	50	55	60	65	70	75	80	85
Area centile		0.02%	0.1%	0.6%	2.3%	6.7%	15.8%	30.8%	50.0%	69.2%	84.2%	93.3%	97.7%	99.3%	99.8%	99.9%
Raw scores Speed-of-reading	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100	105

FIGURE 11. **Normal Curve**

This curve always has the same relation between X and Y values. Thus, if the height at the mean is 100, then the height at + or − one SD is 60. See Table 13 for other values. The proportion of the area which lies between the curve, the base line, and any two perpendiculars can be calculated with great accuracy.

V. OGIVE CURVE

Another useful way of picturing a distribution of scores shows the number of persons whose scores fall above or below a score (see Figure 12). This is known as a *cumulative frequency curve,* or *ogive curve.* The cumulative frequency, shown in Table 9, Column 5, is secured by adding the frequency of each class interval to the sum of the frequencies of all the intervals below it. An ogive curve is made by laying off the scores on the base line, or abscissa, and the total number or per cent of cases on the ordinate, or vertical line. The number of cases in the lowest class interval is indicated by a dot over the score which is the upper limit of the lowest class interval. The upper limit is used to indicate the number of persons who fell below this score. The number of persons in both the first and second class intervals are added together and indicated by a point over the upper limit of the second class interval. Similar procedures are used for the other intervals. Each point on the curve then represents the number of persons whose scores fell below a

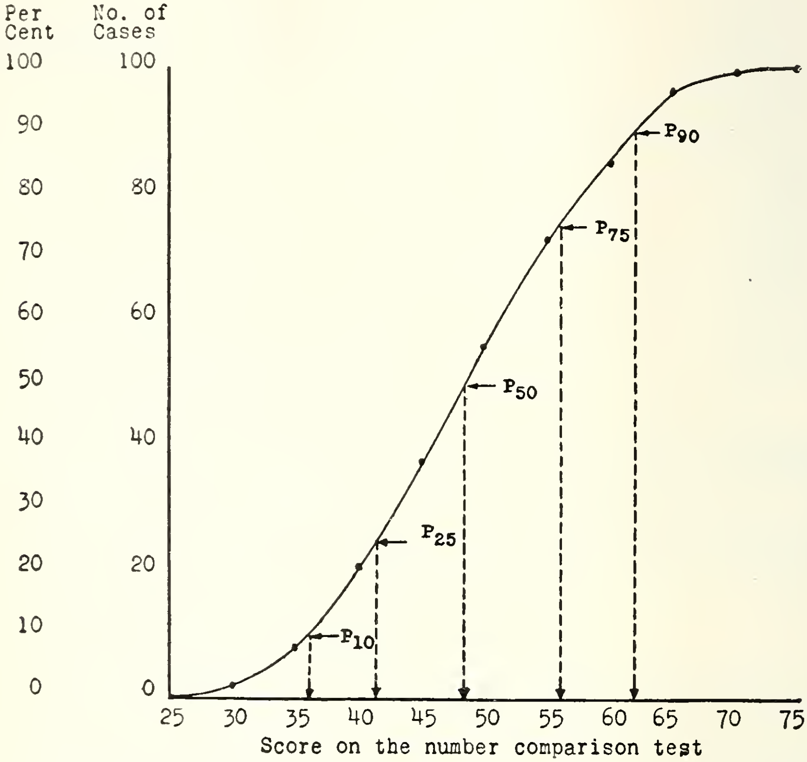


FIGURE 12. **Ogive or Cumulative Frequency Curve of the Scores in Table 8**

particular score. Of course the number falling above any score can also be quickly found from this chart.

VI. RANKS, CENTILES, QUARTILES, DECILES, AND LETTER GRADES

One of the simplest ways of describing a person's position in a group is to indicate his *rank* when the members of the group are arranged in order of size of scores. Ranks are valuable for indicating how many persons secured scores that are above that of a given person. This is all that one needs to know if one is considering test results in situations such as assigning a definite number of scholarships or filling a definite number of positions. However, when knowledge of relative excellence is desirable, as for a wider distribution of grades or other awards, it can be secured by the use of *centiles.* The centile shows the proportion of the group which fell below a given score. Thus, a centile of 75 would mean that one did better than 75 per cent of the group, and not as well as the highest 25 per cent. Centiles may be read directly from an ogive curve such as that shown in Figure 12, or they may be calculated, as in Table 9. Centiles, also called *percentiles,* have as their symbol P, with a subscript to indicate the particular centile. To calculate the 75th centile score, called P_{75}, the score of the 75th person[1] from the bottom must be found, since there are just 100 persons in this group.[2]

The cumulative frequencies seen in Table 9, Column 5, show that 73 scores fell below the class interval 55 to 59.99. Hence, two cases in this interval are needed to reach the 75th in the group. It is customary to calculate this score in all large groups by *interpolation.* Interpolation is the process of estimating a value between two given values by the use of a ratio. In this case, the given values are the limits of the class interval, and the ratio is 2/12, since two of the twelve cases in this class interval are needed. Therefore P_{75} is found by adding 2/12 of the size of the class interval, 5, to the lower limit of the class interval.

---

[1] This calculation would be more precise if the score were found which lies halfway between the 75th and 76th person. For usual work this refinement is not considered necessary.

[2] If the number were 160, the P_{75} would be the score of the 120th person, $(160 \times .75 = 120)$.

$$P_{75} = 55 + (2/12)5 = 55.83$$

Centiles are useful indicators of a person's position in his group, since nearly everyone with a sixth-grade education can understand them. They are probably the most frequently used scores for adults. Table 10 allows raw scores to be changed rapidly to centiles by reading the centile value from the top of the column. For instance, if a boy in the sixth grade received a score of 34, his centile rank would be 93, and it further appears that his score is nearly as high as that of the average ninth-grade student. This table was prepared by securing scores for fairly large groups in each grade and then finding the scores which corresponded to each of the centiles. Such tables are now available for a very large number of standard tests.

Another fairly common way of indicating a person's position in a group is to say in which quarter of the group his score falls. The dividing lines between quarters are called *quartiles,* and these are the 25th, 50th, and 75th centiles.

In some classifications, letter-grade systems are used. In these a definite proportion of a group is arbitrarily assigned a particular letter. Table 11 shows three letter gradings that are in use. From this it appears that in the United States Army examinations a person was given a grade of A if he fell in the highest 4 per cent of the examinees. On the Strong Interest Blank a rating of A is assigned to a person when his score falls among the highest 75 per cent of scores of an occupational group. On a distribution of grades from an elementary psychology class, a rating of A was assigned to the highest 17 per cent of the group. The usefulness of letter ratings is limited, therefore, by the variation in their

TABLE 10

Conversion Table for Michigan Speed of Reading Test

Forms 1 and 2

Items Correct
Detroit Sample

1937 Revision (7 minutes)
N = 3302

Grade	*Age in School*	*Raw Score*	*Standard Score* 30	35	40	45	50	55	60	65	70	*Mean %*
			Centile Rank 2	7	16	30	50	70	84	93	98	*Correct*
Senior	21–0	____	41	45	49	53	57	61	65	69	73	98
Junior	20–1	____	38	42	46	50	54	58	62	66	70	97
Sophomore	19–1	____	36	40	44	48	52	58	60	64	68	95
Freshman	18–1	____	31	35	40	44	49	54	58	62	66	94
12.0	17–2	____	30	34	38	42	46	50	54	58	62	96
11.0	16–3	____	27	31	35	39	43	47	51	55	59	96
10.0	15–3	____	23	27	31	35	40	45	49	54	58	96
9.0	14–5	____	18	22	27	31	36	41	45	49	54	95
8.0	13–6	____	16	20	24	28	32	36	40	44	48	94
7.0	12–7	____	14	17	21	24	28	32	35	39	42	91
6.0	11–7	____	9	12	16	19	23	27	30	34	37	87
5.0	10–6	____	3	6	10	13	17	21	24	28	31	80
4.0	9–5	____	0	2	5	8	11	14	17	20	23	72
3.0	8–4	____	0	0	0	0	3	6	9	12	15	40
		Equivalent I.Q.'s	70	78	85	93	100	107	115	122	130	
		Letter Grades	E	D		C		B		A		

NOTE — (All numbers refer to lower limits of class intervals)
(Greene, 1937)

TABLE 11

The Assignment of Letter Ratings to Proportions of a Group

	1.	2.	3.
A	4.09%	75%	17%
B	8.82	25	25
C +	16.69	0	
C	26.78	0	33
C –	21.86	0	
D	14.38	0	18
E	7.38	0	7

1. United States Army, June, 1918. Memoir Nat. Acad. Sciences. 1921, XV, p. 421.
2. Strong's Interest Blank (1931).
3. A distribution of grades from an elementary psychology class.

meanings. Within certain restricted systems, letter grades allow a quick and convenient comparison of individuals.

The indicators of position in a group which have been described are useful in many situations, but they are not usually considered to be as accurate as the dimensions of a group to be described, and they also fail to provide a rapid comparison of two or more groups.

VII. DIMENSIONS OF A GROUP OF SCORES

Frequency tables and their corresponding curves have three aspects or dimensions: central tendency, dispersion of scores, and shape. A central tendency is a single score near the center of a group which may be used to represent the standing of the whole group. Dispersions show the range of scores found in various portions of a group. The shape of a curve indicates its symmetry or skewness, and its irregularities. These dimensions can be described by numerical indicators which are very useful for comparing groups, and persons within the groups.

A. Central Tendencies

There are three common measures of the central tendency of a group: the *arithmetic mean,* the *mode,* and the *median.* All of these indicators of central tendency are called *averages.*

1. Mean. In this book "the mean" refers to the arithmetic mean unless otherwise indicated. It represents a point of balance which would be found if all scores in a group were assigned the same weight and then arranged along a horizontal beam according to class intervals. The histogram in Figure 10 shows a picture of such an arrangement. If this histogram were a pile of bricks on a straight beam,[3] then it would be possible to find a point of balance at which the bricks on one side of the point would equal in weight the bricks on the other side. This concept of balance

---

[3] To make a better illustration, the beam itself should have no weight at all.

runs through many treatments of psychological material. It depends upon two basic assumptions. One is that the psychological scale is comparable with a *linear* scale, such as distance in a straight line. The other is that members of a group can be considered to be a number of equal, unrelated weights; that is, their measured qualities are unaffected by their position. Both of these assumptions have been challenged for special situations, but they seem to be appropriate in many test situations, and the mean is probably the most commonly used central tendency.

One way to find a mean is to add the scores of all the persons in the group and to divide this total by the number of persons in the group. The mean for the scores in Table 8 by this method is 47.96. If a large number of persons have been measured, an adding machine is a great timesaver.

There is also a short-cut method to find a mean by first guessing at the probable mean score, and then making the necessary corrections. In Table 12 the guessed mean is arbitrarily taken as the midpoint of the class interval, 45–49.9, which seems from inspection to be nearest the middle of the distribution. The deviations, or number of steps, of all class intervals from this class interval are placed in Column 3, and the frequency of class interval is multiplied by its respective deviation, as in Column 4. The deviations above the guessed mean are marked plus, and those below, minus. The total minus deviations are subtracted from the plus deviations $(90 - 69 = +21)$ to find the amount and direction of the correction for guessing. The result, +21 deviations, divided by the total number of cases, is the mean correction. This amount is multiplied by the size of the class interval to make its units of the same denomination as the units of the raw score. Finally, this correction is added to the guessed mean, 47.5.[4] The result is 48.55 which is taken as the mean of the distribution. Table 12 also gives a general equation with letters substituted for the words and phrases used here. Such equations are a great convenience, for the letters or algebraic symbols are more quickly written and read than the words which they represent. Since these symbols are widely used, it is advisable to learn them.

---

[4] This location of the midpoint at 47.5 depends upon the assumption that a score of 45 includes 45.0–45.9, and means that a person has completed 45 items, but not 46.

TABLE 12

Calculation of Mean and Standard Deviation of Scores in Table 8

N = 100

1 Scores	2 Frequencies	3 Deviations	4 Frequency of Deviations	5 Squared Deviations
X	f	x'	fx'	$f(x')^2$
70	1	5	5	25
65	2	4	8	32
60	12	3	36	108
55	12	2	24	48
50	17	1	17	17
45	18	0	0	0
40	17	− 1	− 17	17
35	13	− 2	− 26	52
30	6	− 3	− 18	54
25	2	− 4	− 8	32
sums (Σ)	100		90	385
			− 69	
			21	

$$M = G.\,M. + \frac{(\Sigma x')i}{N}$$

$$= 47.5 + \frac{(21)5}{100} = 47.5 + 1.05$$

$$= 48.55 \text{ or } 48.6$$

$$SD = i\sqrt{\frac{\Sigma(x')^2}{N} - \left(\frac{\Sigma x'}{N}\right)^2}$$

$$= 5\sqrt{3.85 - (.21)^2} = 53.81$$

$$= 5(1.95) = 9.75 = 9.8$$

Legend: M = Mean
G. M. = Guessed Mean
Σ = Sum of
N = Number of Cases
X = Raw Scores
x′ = Deviation from Guessed Mean
i = Class Interval Size
SD *or* σ = Standard Deviation
f = Frequency of Cases in a Class Interval

2. MODE. The *mode* of a group is defined as that score or class interval which has the largest frequency. This is found by making a frequency distribution and inspecting it. Table 12 and Figure 10 show one mode which is at 17.5, the midpoint of class interval with largest frequency, 18 persons.

In some cases a frequency curve shows two or more modes. Two class intervals which have large frequencies may be separated by class intervals with smaller frequencies, as in Figure 13. Sometimes these bimodal, or two-modal, curves result from combining data from what would otherwise form two separate unimodal curves. This is apparently the case in Figure 13. Occasionally bimodal curves are the result of poor measurement techniques. As bimodal curves are rather rare in the measurement of human skills under normal conditions, they need to be scrutinized carefully, and the data on which they are based should be analyzed to find the cause.

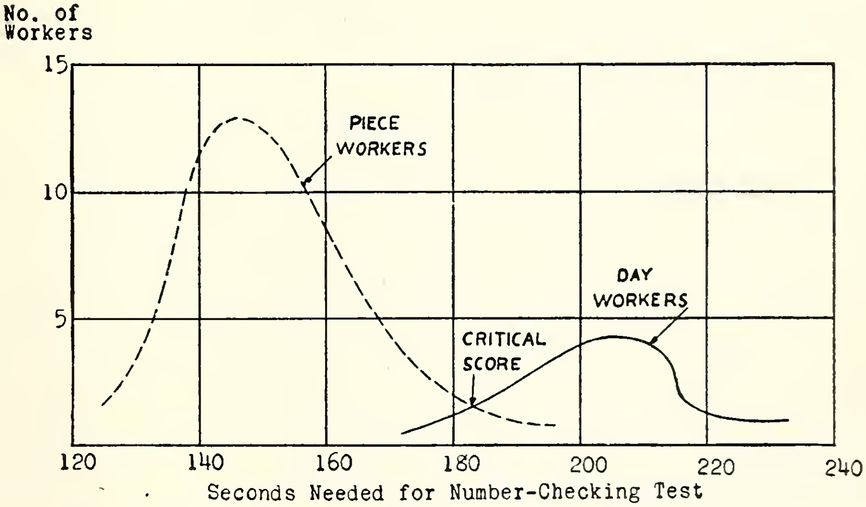


FIGURE 13. **Bimodal Curve, Critical Score**

(From Link, 1918. Reproduced by permission of the Macmillan Co.)

3. MEDIAN. The *median* is defined as the middle score in a group when the scores are arranged according to size

or P_{50}. The median is found by an interpolation of the scores in the class interval which contains the middle score. If we take the middle score as that of the 50th[5] person in a group of 100, or the 50th centile, then in Table 9, Column 5, it appears that this centile falls in the class interval 45–49.9. Thirty-eight persons are below this class interval; thus the twelfth person from the bottom of this class interval is the 50th person in the group. His approximate score can be calculated by interpolation when it is assumed that the 18 scores in this class interval are evenly distributed over the whole interval. When the interval is small compared to the number of cases falling in this interval, this assumption does not introduce an appreciable error. From this assumption 12/18 of the size of the class interval, 5, must be added to its lower limit, 45, to secure the score of the 50th person. In this case the median is 48.3.

$$\text{Median} = 45 + \left(\tfrac{12}{18}\right)5 = 48.3$$

4. THE USES OF CENTRAL TENDENCIES. The question of which central tendency is the best often arises. The answer depends upon the distribution of the scores and their intended use. In a normal curve the three central tendencies are identical, but in an irregular or a skewed curve they will differ (see Figure 14). A curve is said to be skewed from the normal when it is not symmetrical. Let us consider a situation in which the most representative score of a group is desirable. Of the three, the mean is most affected by the extreme scores of a group, so when the extreme scores are thought to be rather insignificant, the median or mode is used. The mean represents a true center of balance in a group, and it is generally preferred when the scores represent very precise measurements. The mode indicates the most

---

[5] For greater accuracy, the 50.5th person's score should be used. When the number of persons is even, the midscore lies midway between two mid-mosts scores. However, this degree of accuracy is usually sacrificed for economy. When universally applied, comparisons of persons or groups are not affected by taking the median as the 50th centile.

frequent score in a group, and it also shows when bimodality occurs.

In addition to its use as the most representative score of a group, a central tendency is also used in calculating various interesting ratios which are described later in this book.

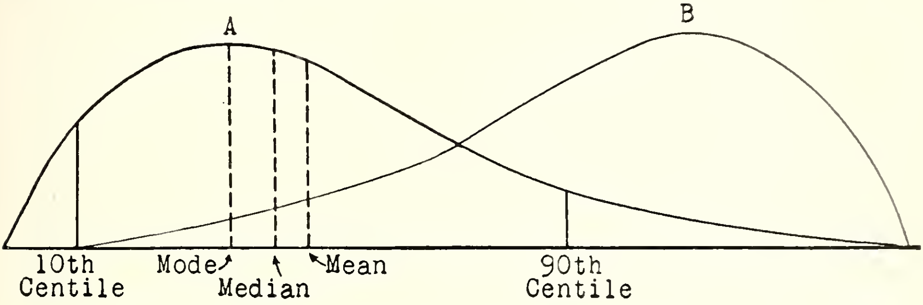


FIGURE 14. **Skewness**

B. Dispersions

The amount of dispersion exhibited by a group is usually indicated by one of four indicators: the total range, the quartile deviation, the probable error, and the standard deviation.[6]

1. Total Range. The total range of scores is the difference between the highest and lowest scores in a group. Its use in securing class intervals has already been described. It is not, however, used commonly for comparing the dispersions of two groups of persons because its size depends upon the two extreme scores. These two scores often vary considerably in several chance samplings, particularly in small groups, so that other indications of dispersion which are less affected by chance are more useful.

2. Quartile Deviation and Probable Error. Both the quartile deviation (Q) and the probable error (PE) are found by taking one-half the range of the middle 50 per cent

---

[6] Another indication of dispersion which is also used occasionally is called the average deviation, AD, or *mean deviation*, MD, or *mean variation*, MV. It is found by summing all deviations from the mean regardless of sign, and dividing this sum by the number of scores. In a normal curve the AD is one-half the range of the middle 57.5 per cent of cases.

of the cases in a group. The Q is used for actual measures, and the PE is applied to hypothetical situations where the probable variation which would occur through chance is desired for purposes of prediction. Q is found by subtracting the 25th centile from the 75th centile, and dividing the remainder by 2.

$$Q = \frac{P_{75} - P_{25}}{2}$$

The calculation of Q is shown in Table 9.

The probable error (PE) is usually calculated by multiplying the standard deviation by .6745. The figure is the ratio of the PE to the SD of all normal curves. Both the standard deviation and PE are used only for normal distribution curves.

3. STANDARD DEVIATION. If one drops perpendiculars from the two inflection points of a normal curve of distribution to the base line, the total area thus bounded includes 68.27 per cent, or roughly two-thirds of the whole area. The distance on the base line from the mean to the foot of such a perpendicular is known as a *standard deviation.* Its usual symbols are a small sigma (σ) or the initials SD. The SD, used as a unit of measurement. can be laid out along the base line a number of times, as in Figure 11. The normal curve extends approximately 2½ SD in each direction from the mean, and the total length of the base line needed to include 98.76 per cent of the cases is 5 SD. The per cent of the area which lies above each hundredth of an SD has been calculated to twelve decimal places. A standard deviation has therefore become a very precise indicator of position in a group. The SD has another advantage: it produces a scale whose steps are equal in the sense that they represent equally-often-noticed differences. Theoretically, the difference between scores of persons who are .5 and .6 SD above the mean can be noticed by a competent group of judges just as often as a difference of .1 can be noticed anywhere in

the scale. This provides a method of scaling scores of psychological phenomena, described in Chapter 22, which is comparable with the best physical scales, since all scales are based on the ability of competent judges to notice certain differences equally often.

4. THE SHAPE OF A CURVE. Before calculating the standard deviation for any group, one must be reasonably sure that his figures fit a normal distribution curve fairly well. Tests for *goodness of fit* include those for *skewness* and *kurtosis.* A curve is said to be skewed from the normal when it is not symmetrical. Kurtosis is the relative flatness of a curve. A curve varies from normal kurtosis when it is relatively more peaked or more flattened than the normal curve.

Skewness (Sk) can be roughly measured by the formula (Garrett, 1937, p. 299) :

$$\mathrm{Sk} = \frac{P_{90} + P_{10}}{2} - P_{50}$$

When Sk is zero, the curve is symmetrical. When Sk is positive, the curve extends farther to the right than to the left of the mean, as in Figure 14A. When Sk is negative, the curve will resemble Figure 14B. From Table 9 the skewness of the distribution is found to be .60.

The kurtosis (Ku) of a curve can be roughly found by the formula (Garrett, 1937, p. 230)

$$\mathrm{Ku} = \frac{Q}{(P_{90} - P_{10})}$$

This formula gives a value of .263 for the normal curve. The distribution in Table 9 shows a kurtosis of .275. The amount by which a curve may deviate from the normal shape in either skewness or kurtosis by chance errors of measurement can be known from formulae in Garrett or other statistical texts. Small variations from the normal curve introduce insignificant errors in practical comparisons.

5. Calculation of a Standard Deviation. Since the curve in Figure 10 has shown nearly normal values for both skewness and kurtosis, the formulas which have been designed for use with a normal curve may be safely applied. The calculation of the standard deviation for the scores of Table 8 is shown in Table 12. It is found by the following steps:

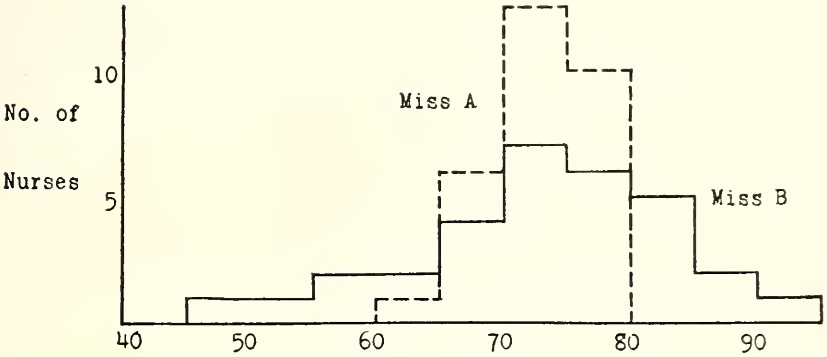
1. The deviations from the guessed mean are squared and added in column 5.
2. The sum of the squared deviations is divided by the number of cases. This result is the mean of the d^2s.
3. The mean of the squared deviations is corrected for the error caused by the use of the guessed mean by subtracting the square of the mean deviation. If the deviations had been measured from the true mean, this step would not have been necessary, but it is nearly always quicker to use a guessed mean and the correction than to find the correct mean and then measure deviations from it.
4. The square root of the mean of the squared deviations is found and multiplied by the class interval. This is the SD.

VIII. USES OF MEASURES OF DISPERSION

Two common uses of measures of dispersion — the comparison of groups and the comparison of individuals in a group — are described below.

A. Group Comparisons

Very often an investigator wishes to indicate which of two groups has a larger dispersion of scores on a test. For instance, in employment or educational work it is important for administrators to know the range of abilities of the persons in various groups with whom they are dealing. When the groups have a normal form of distribution, either the quartile deviation or the standard deviation is a very convenient measure of dispersion. When groups of 50 or less are compared, the Q is used because it is less affected by extreme scores than is the SD. Figure 15 shows two groups of service ratings which are compared in this fashion.



Both supervisors gave the same median score for the group, 72.6. Miss A assigned no very high or low grades, hence their quartile range (Q) is 3.1. Miss B used a much wider range of scores so that their Q is 7.7.

FIGURE 15. **Differences in Dispersion**

B. Individual Comparisons

A very frequent use of measures of deviation is the indication of a person's position in a known group of persons. In such cases the standard deviation is universally used rather than any other measure. A person's raw score may be changed to a *standard deviation score* by subtracting the mean (M) of the group from the person's score (X), and dividing the remainder by the standard deviation (σ) of the group. Thus, if John scored 39 in Table 8, he would have a standard deviation score of –.99 since his score would be –.99 standard deviations below the mean.

$$\text{SD score} = \frac{X - M}{\sigma} = \frac{39 - 48.6}{9.8} = \frac{-9.6}{9.8} = -.99$$

In order to eliminate decimal points and minus signs which are somewhat troublesome, *standard scores* are often used. Standard deviations and equivalent standard scores are shown in Figure 11 and Table 13. In standard scores, the mean of a group is placed arbitrarily at 50, and one standard deviation is given the value of 10. Thus, John whose raw score is 39 would have a standard score of 40.1.

$$\text{Standard Score} = \frac{(X - M)10}{\sigma} + 50 = \frac{(39 - 48.6)10}{9.8} + 50$$
$$= -9.9 + 50 = 40.1$$

TABLE 13

Corresponding Values of Standard Deviations, Standard Scores, and Centiles of the Normal Probability Curve

(1) *Standard Deviation*	(2) *Standard Score*	(3) *Centiles*	(1) *Standard Deviation*	(2) *Standard Score*	(3) *Centiles*
+ 3.0	80	99.86	− 0.1	49	46.02
2.9	79	99.81	− 0.2	48	42.07
2.8	78	99.74	− 0.3	47	38.21
2.7	77	99.65	− 0.4	46	34.46
2.6	76	99.53	− 0.5	45	30.85
2.5	75	99.38	− 0.6	44	27.43
2.4	74	99.18	− 0.7	43	24.20
2.3	73	98.93	− 0.8	42	21.19
2.2	72	98.61	− 0.9	41	18.41
2.1	71	98.21	− 1.0	40	15.87
2.0	70	97.72	− 1.1	39	13.57
1.9	69	97.13	− 1.2	38	11.51
1.8	68	96.41	− 1.3	37	9.86
1.7	67	95.54	− 1.4	36	8.08
1.6	66	94.52	− 1.5	35	6.68
1.5	65	93.32	− 1.6	34	5.48
1.4	64	91.92	− 1.7	33	4.46
1.3	63	90.32	− 1.8	32	3.59
1.2	62	88.49	− 1.9	31	2.87
1.1	61	86.43	− 2.0	30	2.28
1.0	60	84.13	− 2.1	29	1.79
0.9	59	81.59	− 2.2	28	1.39
0.8	58	78.81	− 2.3	27	1.07
0.7	57	75.80	− 2.4	26	.82
0.6	56	72.51	− 2.5	25	.62
0.5	55	69.15	− 2.6	24	.47
0.4	54	65.54	− 2.7	23	.35
0.3	53	61.79	− 2.8	22	.26
0.2	52	57.93	− 2.9	21	.19
0.1	51	53.98	− 3.0	20	.135
0.0	50	50.00			

(Arranged from Pearson, 1914. By permission of Editor, *Biometrika* publications.)

C. Standard Score versus Centiles

Standard scores have an advantage not enjoyed by centiles, in that scores may be added and subtracted without the fear of introducing errors due to inequality of scaling. This can be understood when one looks at the normal frequency curve, Figure 11. It is evident that the distance on the base line between the 50th and 69th centile is the same as the distance between the 93rd and 97th centiles. This is due to the fact that in a normal curve the frequency of scores is greater toward the center of the distribution. If a person's scores on two tests are averaged, using centiles, an error is introduced. For example, if Frank made standard scores of 50 and 70 on two tests, the mean of his standard scores is 60, $\frac{(50 + 70)}{2}$. These standard scores correspond to the 50th and 97.72 centiles on Table 13. The mean of these centiles is 73.8, $\frac{(50 + 97.72)}{2}$, which corresponds to a standard score of only 56.4 instead of the correct standard score of 60. The use of centiles has introduced an error in this case of 3.6 points.

D. Profiles

One of the best indications of a person's ability is a graphic record of his scores on several tests that are considered to be important and somewhat independent. Such a graphic record, called a *profile,* is constructed by plotting a person's scores in a profile chart. A profile chart is made by laying off on graph paper a line which represents the mean of a group and other parallel lines which represent units of dispersion, such as standard scores. A profile is by far the most interesting and valuable way to represent results of tests. for it gives a person a clear picture of his strong and weak subjects. For the greatest convenience the raw scores are placed at the proper intervals for each test in order that the person may find his position in a group by simply placing dots over the numbers which represent his scores. Lines connecting these dots help to make a quick comparison. In Figure 16 mean score profiles are shown for four occupational groups on eight different tests based on standard scores of a large adult population.

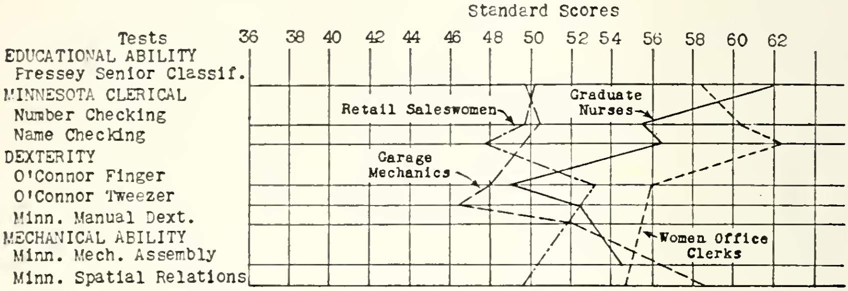


FIGURE 16. **Occupational Profiles from Standard Tests**

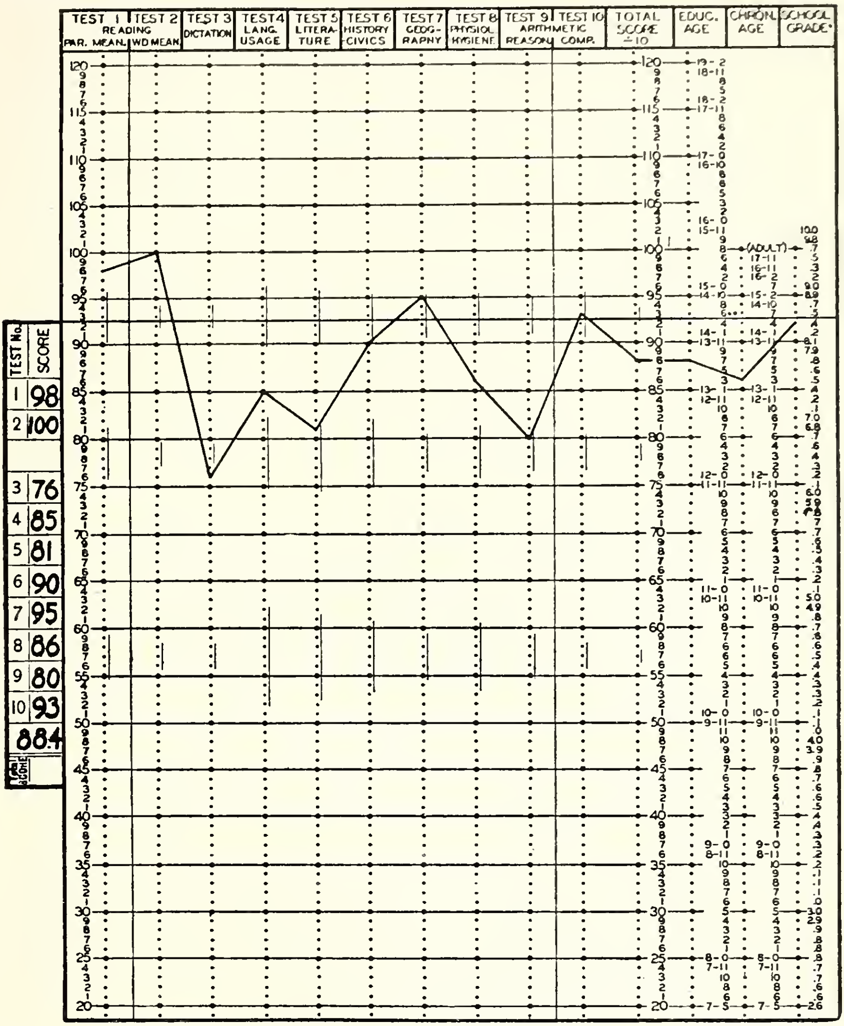
(Adapted from Dvorak, 1935. By permission of the University of Minnesota Press.)

Similar profiles may be constructed from age scales in which the scores in each test have been given chronological age equivalents. Thus, Figure 17 shows a profile from a Stanford Achievement Test. This profile has the advantage of showing scores for average age groups and grades in school. It also has the advantage of being composed of steps which are probably as equivalent as the steps in the standard deviation scales, because age differences become smaller with advancing age (see scaled growth curve, Chapter 21).

E. To Combine a Person's Score on Several Tests

Occasionally one wishes to have a single score which will represent a person's ability on several different tests. The selection of the best way to combine scores depends upon what use is to be made of the combination. One way to combine test results is to average a person's standard scores. In some cases, where very few tests are involved, the median is preferred since it is less sensitive to extreme scores than is the mean.

EDUCATIONAL PROFILE CHART: NEW STANFORD ACHIEVEMENT TEST, ADVANCED EXAMINATION



* Grade defined as in Table 2 of the *Directions for Administering*.
** Educational Ages above this point are extrapolated values.
See *Guide for Interpreting* for explanation of vertical lines.

FIGURE 17. **Educational Achievement Profile of Eleanor Brown**

(Kelly, Ruch, and Terman, 1929. By permission of the World Book Co.)

F. Weighted Scores

In some instances it is considered desirable to give more weight to the scores in one test than to the scores in another test. Methods of weighting the scores are of two kinds. One method is based on a desire to give more credit for a harder task or a task that is considered more important than another. This is generally a rather arbitrary matter such as is found when one score is given four times as much credit as another. Thus, in the United States Army Alpha examination the number of items correct on the arithmetic test was multiplied by four, whereas the scores on several other tests were simply the number of items correct.

The other method uses a statistical analysis which shows how well various scores predict a certain criterion. In order to make the best prediction, scores are given various weights. Methods of combining scores in a way to make the best possible prediction of a particular criterion are to be found in standard statistical texts.

An able discussion of the effects of weighting various standard scores upon their combined results was given by Scates and Fauntleroy (1938). They pointed out that the differences between various weightings depend upon:

1. The number of scores which are combined
2. The uniqueness of the various scores
3. The shape of the distribution curve of each score
4. Whether the weights are constant or variable

The evaluation of the weighting of scores is an intricate process which depends upon both practical and theoretical considerations. It is discussed later in the book in connection with several particular tests in Part II.

4

MEASURES OF RELATIONSHIP

I. PREDICTIONS OF PROBABLE SCORES FROM KNOWN SCORES

To a great extent, psychological measurement is designed to render an accurate prediction of probable quality and quantity of individual development. Predictions are based upon the assumption that an individual will develop as persons known to be similar have developed in the past. They are only possible on the basis of knowledge of past relationships in similar cases. They may take the form of a statement of the probability that a person will have a certain score in one variable when he has a particular score in another variable. A variable is defined as any aspect in which members of a group differ from one another, such as height, location, or various abilities and attitudes.

The technique for ascertaining such probabilities is relatively simple. Much credit for developing this technique goes to Sir Francis Galton (1886) in connection with his studies of the inheritance of genius. His work is still among the best in this field. Starting with sheets of graph paper and some pins, he devised a method of measuring relationships among variables, now called the product-moment method of correlation. This method was perfected by Pearson (1904).

A. Scatter Diagrams

In finding relationships between parents and children, Galton arranged his data on graph paper, as in Figure 18, so that the heights of adult children were indicated on the base line, the X-scale, and the heights of midparents (average of the two parents) on a vertical line, the Y-scale. For each parent-child pair he placed a pin to show the scores of both the midparent and the child. Thus, a midparent's score of 71 and a child's score of 64 would be represented by a pin in the square marked A, which lies in the column containing the child's score and the row containing the adult's score. The pins in each cell, or square, were then added as in Table 14. Such charts are called scatter diagrams. They allow one to see at a glance the distribution of relationship.

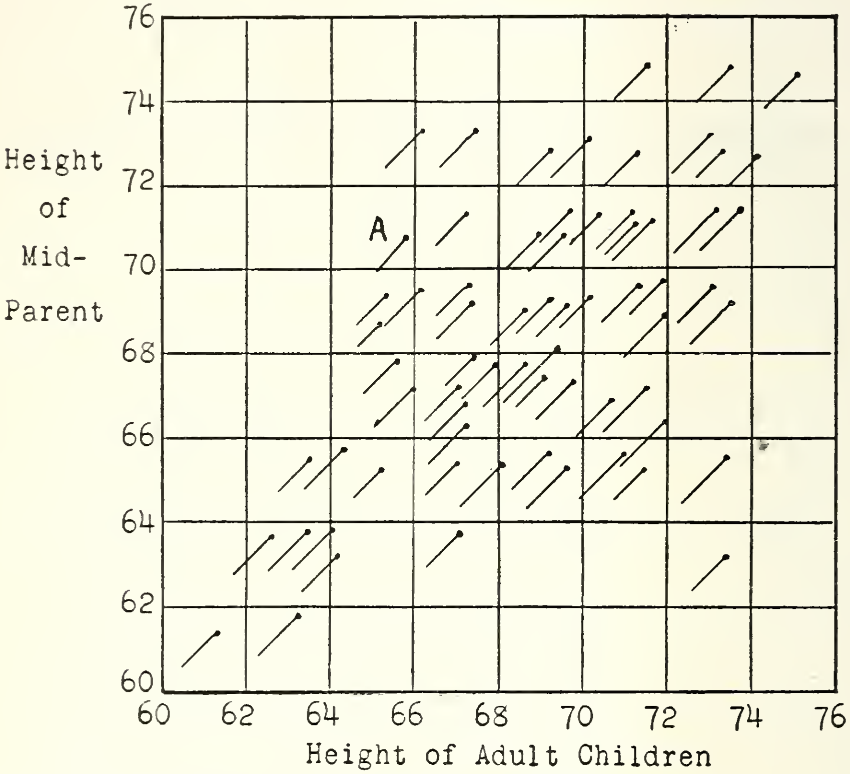


FIGURE 18. **Scatter Diagram with Pins**

TABLE 14

Correlation Table

Adult Children — i = 1 inch

Parents

	61	62	63	64	65	66	67	68	69	70	71	72	73	74		y'	f	fy'	f(y')²	x'y' +	x'y' −
73												1	3			5	4	20	100	95	
72								1	2	1	2	7	2	4		4	19	76	304	288	
71					1	3	4	3	5	10	4	9	2	2		3	43	129	387	285	39
70	1		1		1	1	3	12	18	14	7	4	3	3		2	68	136	272	232	40
69			1	16	4	17	27	20	33	25	20	11	4	5		1	183	183	183	237	142
68	1		7	11	16	25	31	34	48	21	18	4	3			0	219	544			
67		3	5	14	15	36	38	28	38	19	11	4				-1	211	211	211	254	125
66		3	3	5	2	17	17	14	13	4						-2	78	156	312	220	42
65	1		9	5	7	11	11	7	7	5	2	1				-3	66	198	594	378	81
64	1	1	4	4	1	5	5		2							-4	23	92	368	268	8
63	1		2	4	1	2	2	1	1							-5	14	70	350	210	5
x'	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7			727			
f	5	7	32	59	48	117	138	120	167	198	64	41	17	14			928	-183	3081	2467	482
fx'	35	42	160	236	144	234	138	989	167	198	192	164	85	84	890	-99	Sum x'	Sum y'	Sum y'²	Sum x y	1985
f(x')²	245	252	800	944	432	468	138		167	396	576	656	425	504		6003	Sum x'2				

Answers

Formula 1. $\text{Mean X} = \text{Guessed Mean} + \frac{\Sigma x'}{N} i = 68.5 + (-.105)1 =$ 68.39

Formula 2. $\text{Mean Y} = \text{Guessed Mean} + \frac{\Sigma y'}{N} i = 68.5 + (-.199)1 =$ 68.30

Formula 3. $\text{SD}x'^2 = \sqrt{\frac{\Sigma x'^2}{N} - \left(\frac{\Sigma x'}{N}\right)^2} = \sqrt{6.47 - .0114} = \sqrt{6.46} = 2.55$ 2.55*

Formula 4. $\text{SD}y'^2 = \sqrt{\frac{\Sigma y'^2}{N} - \left(\frac{\Sigma y'}{N}\right)^2} = \sqrt{3.33} - .039 = \sqrt{3.30} = 1.82$ 1.82*

Formula 5.
$$r = \frac{\frac{\Sigma x'y'}{N} - \left(\frac{\Sigma x'}{N} \cdot \frac{\Sigma y'}{N}\right)}{\text{SD}_x \cdot \text{SD}_y} = \frac{2.11 - (-.105)(-.198)}{2.55 \quad 1.82}$$
$$= \frac{2.09}{4.63}$$
.451

Formula 6. $\text{PE}_r = .6745 \frac{1 - r^2}{\sqrt{N}} = .6745 \frac{1 - .2034}{\sqrt{928}} = .6745 \frac{.7966}{30.46} =$ .014

Formula 7. $K_r = \sqrt{1 - r^2} =$ .898

* Must be multiplied by i if SD_x and SD_y are desired.

(From data in Galton, 1886.)

B. Regression Lines

In order to predict the most probable height of a child from midparents of a particular height, say 69 inches, Galton assumed that the most frequent height would be the same as the median height of children in the particular row of 69-inch parents. A glance at Table 14 shows this median height to be a trifle less than 69 inches.

The general relationship between parents' and children's heights will be indicated by a line drawn through the medians of the vertical rows. Such a line is called a *regression* line of children on parents. A similar line drawn through the midpoints of the horizontal columns is called the regression line of parents on children. When both of these lines happen to be straight, one of them alone is enough to show both relationships. Table 14 exhibits regression lines that are slightly curved and broken. Straight lines could be drawn in such a way, however, that they would pass very near to the median points. Since this is the case, the regression lines can be treated as if they were straight.[1]

Upon inspection of Table 14 it will be seen that tall midparents, who were in the class interval labeled 70, had children whose median height was lower, 69.9 inches. Short midparents, in class interval 63, had children whose median height was higher, 65.0 inches. From such measures Galton generalized his famous law of filial regression, namely, that children tend to be more like the central tendency of the group than were their parents.

The term regression line is now applied to all lines which indicate the relation between the class intervals on one scale and the corresponding mean or median scores on another scale. This relation can be found and given a quantitative value for the group as a whole, when the regression line is straight, by measuring its slope from the base line.

---

[1] The line of best fit, that is, the line which has the smallest total deviation from all the dots, can be calculated. Its slope is given by the product-moment correlation coefficient described below.

C. Slope of Regression Lines; Correlation Coefficients

The slope of a straight regression line from the base line can be indicated by the ratio between distances from the horizontal and vertical scales or axes, and any point on the regression line except zero. Thus, in Figure 19, Diagram 3, the point "P" is 3 units from the horizontal axis or base line, and 5 units from the vertical axis. The ratio between these two lines ($3/5 = .60$) will be the same as the ratio of two similar lines measured from another point on the same regression line. The slope, in this case .60, is known in trigonometry as the *tangent* of the angle between the regression line and the base. Since tangents have been calculated very carefully for many angles, the tangent is a convenient figure to use for the slope of a line. When the horizontal and vertical scales are changed to standard-deviation units, variations in raw scores are eliminated from a scatter diagram. In this case the tangent of the regression angle is the *correlation coefficient.* The correlation coefficient thus shows the slope of a regression line when the two tests are scaled alike. For a 45-degree angle the tangent and the correlation coefficient are 1.00. If the regression angle approaches zero, the correlation coefficient approaches zero. A zero indicates that the relationship between the variables compared is merely chance. A *chance relationship* means that a person who makes a particular score in one test has the same probability as anyone else in the group of making any one of the scores in the other test. Figure 19 shows regression lines and correlation coefficients for six typical situations. Diagrams 1 to 4 illustrate elliptical boundaries and regular distributions which probably indicate that both variables are normally distributed. If the tally marks show large gaps as in Diagram 5, it is clear that the variables are irregularly distributed. If the marks show a fan-shaped distribution as in Diagram 6, the correspondence of scores between the two variables is evidently much closer at the lower end than at the higher end of the distribution.

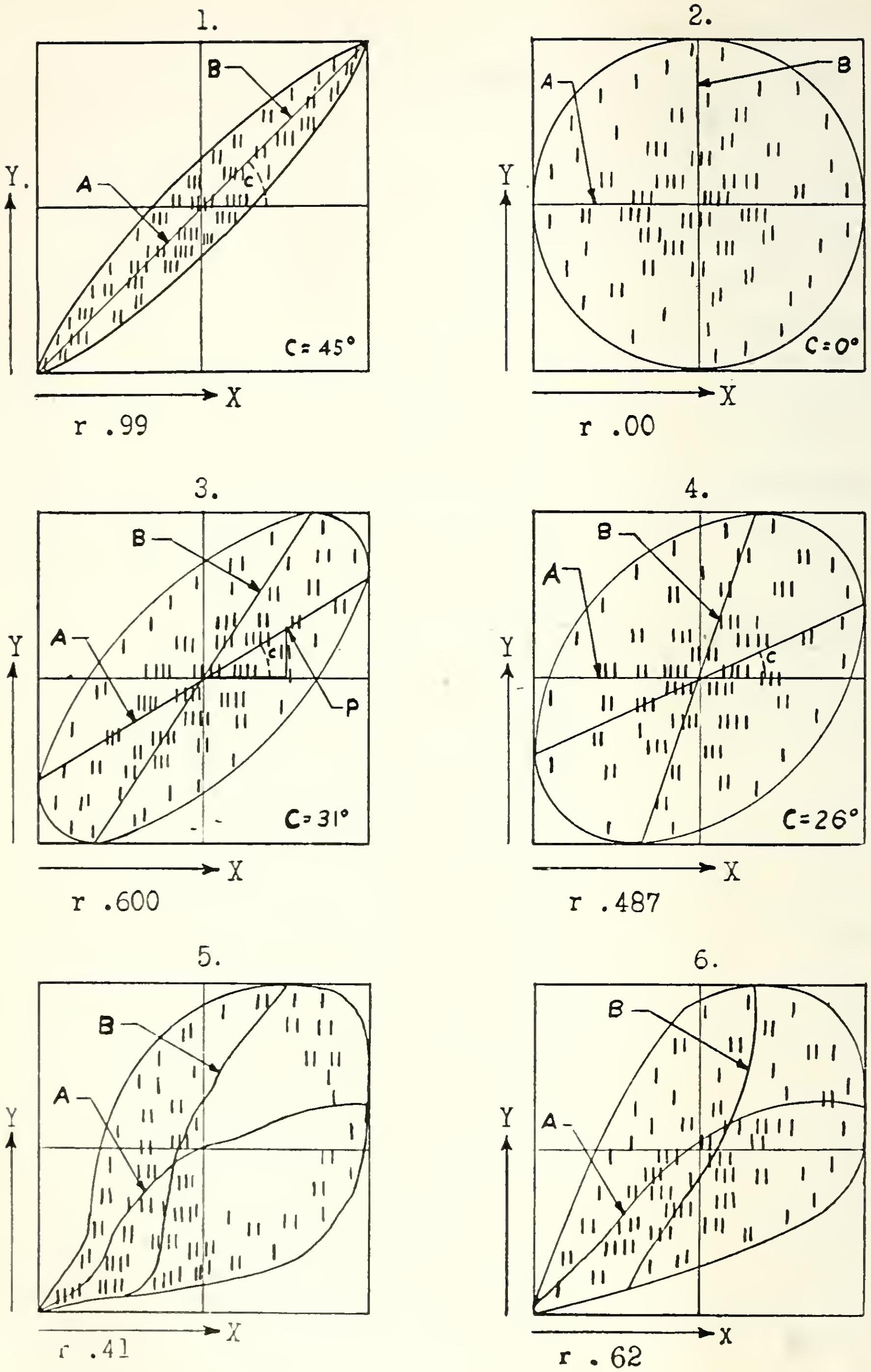


FIGURE 19. Typical Scatter Diagrams and Regression Lines

D. Product-Moment Correlation Coefficients

Since the process of making scatter diagrams and drawing regression lines is laborious, shorter methods have been devised for calculating relationships. One of the best and most commonly used is called the *product-moment correlation* method. It is designed for use when both tests show fairly normal distributions, when the regression lines are approximately straight, and when large numbers of individuals are involved.

The product-moment correlation method makes use of the fact that if all persons have exactly the same relative position on two tests the correlation is perfect. But the more they differ in relative position, the more the correlation is lowered. Since relative position is well indicated by SD scores, one may proceed to find the difference (d) between the two SD scores for each person, taking account of their signs. The following formula will then yield a product-moment correlation, r, when n is the number of individuals:

$$r = 1 - \frac{\Sigma d^2}{2n}$$

Product-moment correlations are, however, usually found, as shown in Table 14, without the labor of computing SD scores. The steps are as follows (Formulas 1 and 2 are used to secure the means and are not essential to the product-moment method):

1. A scatter diagram is constructed.
2. A *Guessed Mean* is located and deviations x′ and y′ are indicated for each variable respectively and summed. These sums are used to correct for guessing.
3. The frequency, f, of each row and column is found.
4. The frequencies are multiplied by the deviations squared, and substituted in Formulas 3 and 4 to secure the standard deviations.
5. The frequency of each cell is multiplied by its x′y′ value. The x′y′ values are the product of the x′ and y′

deviations for each cell. The x′y′ values are summed in the two right-hand columns and substituted in Formula 5 to secure the correlation coefficient.

E. Rank-Order Correlation Coefficients

When a group is small, correlation coefficients are often secured by the rank-order method because it is not so much affected by extreme scores as the product-moment coefficient. Table 15 shows how a correlation of .564 was found between clerical test scores and a supervisor's ranking of ten persons. First, the scores are changed to ranks. Then, the differences (D) between each person's two ranks are found and squared. Lastly, the sum of the squared differences is substituted in the formula:

TABLE 15

Rank-Order Correlation Method

INDIVIDUAL	CLERICAL TEST		SUPERVISOR'S RANKING	D	D²
	Score	*Rank*			
Metcalf, C.	76	1	6	5	25
Adams, J.	75	2	3	1	1
Bushee, L.	43	3	1	2	4
Jones, F. C.	42	4	2	2	4
Coover, L.	41	5	5	0	
Cook, F.	36	6.5	8.5	2	4
Tostlebee, F.	36	6.5	8.5	2	4
Bahrman, K.	27	9	10	1	1
Hewitt, T. S.	27	9	7	2	4
Shumaker, L.	27	9	4	5	25
					72

$$N = 10$$

$$\text{Rho} = \rho = 1 - \frac{6\,\Sigma D^2}{n(n^2 -)}$$

$$= 1 - \frac{432}{10 \times 99} = 1 - .436$$

$$= .564$$

$$\text{Rho} = \rho = 1 - \frac{6\,\Sigma D^2}{n(n^2 - 1)}$$

Rho will be 1.00 and thus show a perfect correlation if the D's are all zero. Rho will be .00 if the ranks have only a chance relationship, and – 1.00 when there is a perfect negative relationship. Rho is nearly always a little smaller than r, a product-moment correlation for the same set of figures.

F. Individual Predictions

Individual predictions of the most probable score on one test when the score on another test is known can be made in two ways: one from a scatter diagram, the other from an equation which represents a regression line.

Prediction from a scatter diagram can be made by finding the most frequent score in Y which corresponds to a particular score in X. Thus, in Table 14, the most probable height of a child whose midparents are in the 63.0 class interval is 65.0 inches.

Prediction from a regression equation avoids the rather lengthy process of making a scatter diagram and calculating the central tendencies of each class interval. Regression equations also give a little better prediction since they are less affected by minor variations of a chance variety than are the central tendencies of class intervals. Regression lines smooth out minor chance deflections. In finding the most probable height of children (X scale) from parents who are 63.5 inches tall (Y scale) a convenient formula is

$$\overline{X} - M_X = r_{XY} \frac{\sigma_X}{\sigma_Y} (Y - M_Y)$$

The formula looks more complicated than it is, for it is really a short statement of familiar terms, as:

$\overline{X}$ is the score to be predicted

M_X (called M sub X) is the mean of all X scores = 68.4 inches

Y is the given score = 63.5 inches

M_Y (M sub Y) is the mean of all Y scores = 68.3 inches

r_{XY} is the correlation coefficient found between X and Y which shows the slope of the regression line = .451

σ_X and σ_Y are the standard deviations of the X and Y distributions (included to equalize the group dispersions)

The most probable height of the child is therefore

$$\begin{aligned}\overline{X} &= r_{XY}\frac{(\sigma_X)}{(\sigma_Y)}(Y - M_Y) + M_X \\ &= [.451\frac{(2.55)}{(1.82)}(63.5 - 68.3)] + 68.4 \\ &= 65.37\end{aligned}$$

G. Standard-Error-of-Estimate

Since the score just found, 65.37, is the most probable score, it is of value to find out how accurate the prediction is. A particular son's height may, of course, vary considerably from the most probable height. The probable amount of this variation can be found by calculating the standard deviation of the estimated score. This is called a *standard-error-of-estimate,* and it is written $\sigma_{X \cdot Y}$ when the variation of a predicted X score for a given value of Y is desired. From a standard-error-of-estimate the range of scores which will include any particular proportion of children in a particular row can be found.

A general standard-error-of-estimate [2] to apply to all probable scores is given by the formula

$$\sigma_{X \cdot Y} = \sigma_X\sqrt{1 - r_{XY}{}^2}$$

When applied to the data in Table 1 it becomes

$$\begin{aligned}\sigma_{X \cdot Y} &= 2.55\sqrt{1 - (.451)^2} \\ &= 2.55\ (.887) \\ &= 2.26\end{aligned}$$

[2] The formula is only approximate, since it gives one standard-error-of-estimate for all columns. From inspection of a scatter diagram one can see that this is not actually the case. The columns near the end have smaller dispersions than those in the middle. The standard-error-of-estimate is the average standard deviation of the columns. It is commonly used for individual predictions.

From this it appears that the middle 68 per cent of the children who have parents of a particular height will fall within heights ranging from 2.26 inches above to 2.26 inches below the most probable score. For instance, when the most probable score of a child is 65.37, the middle 68 per cent of the children will range from 63.11 to 67.63 inches in height.

We can also predict what are the chances in 100 that a child of parents of a certain height will attain any given height, by finding the proportion of all the children of parents of this height who have reached the given height. This is done by finding the difference between the most probable height of children and the given height, and dividing this difference by the standard-error-of-estimate. The result of this division will be the number of standard deviations from the most probable height, and this can easily be translated from Table 13 into percentages. Suppose we wish to know what the chances are that Frank, whose parents average 63.5 inches, will be 62 inches or less in height. We have just found that his most probable height will be 65.37 inches and the standard-error-of-estimate is 2.26 inches. The difference, $65.37 - 62$, divided by the standard-error-of-estimate gives 1.49. Table 13 shows that -1.49 standard deviations is equivalent to the 6.9 centile. We may, therefore, conclude that Frank has only about 7 chances in 100 of being as short as 62 inches. These calculations may seem a bit complicated at first, but, after working out several examples, the routine becomes easy. In practice, it is often desirable to make such predictions.

H. Standard-Error-of-a-Score

The same logic has been applied to prediction of the probable variation of any single score on a test. It has been shown that if two tests correlate highly, the standard-error-of-estimate of a score predicted from one test to the other will be very small. Similarly, if a test has a high self-correlation on two trials, the standard-error-of-estimate of a score predicted from one trial to the other will be small. Since in this case the same test has been repeated, some authors conclude that the tendency for a single score to vary by chance from its true value can be shown by the formula for a standard-error-of-a-score. It is expressed thus:

$$\text{Standard-error-of-a-score} = \sigma_{\infty} = \sigma_1\sqrt{1 - r_{12}}$$

Here ∞ indicates the theoretical standard deviation of an infinite number of obtained scores from the most probable true score; σ_1 is the standard deviation of the actual scores in Trial 1 of the test; and r_{12} is the correlation between Trial 1 and Trial 2.

This formula may be applied only when the means and sigmas of the two trials are nearly the same. The standard-error-of-a-score would not show the true tendency toward variation if the scores or their dispersions become larger or smaller with repetitions.

The standard-error-of-a-score is often a more practical indication of the predictive value of a test than the correlation between trials, as it shows the amount of variation which may be expected. In a revision of the Binet Scale, Terman and Merrill (1937) report that IQ's from 90 to 109 have standard errors of 4.51 when the estimated correlation between two forms of the test was .924. This means that the chances are 2 out of 3 that the true IQ is not higher or lower than the obtained IQ by more than 4.51 points when the obtained IQ is between 109 and 90.

II. APPLICATIONS OF CORRELATION TECHNIQUES

The use of correlation techniques has spread remarkably during the last twenty years. Every field of psychological investigation now makes use of them in situations where relationships are to be appraised. A complete list of such situations would be very long but the main applications would include:

A. Related Persons

Pairs of persons, who are related in some way, are measured and the score of one member of a pair is plotted against the score of the other member of the pair. This method is of prime importance in studies of inheritance where familial resemblances are studied. Studies of pairs of identical twins, of mothers and daughters, of fathers and sons, of siblings, of cousins, and of grandparents and grandchildren have shown typical relationships by correlation techniques (see Chapter 24).

B. Related Scores

When a group of persons have been appraised in two situations, the scores in one situation may be plotted against the scores in the other. Literally thousands of studies of this sort have been made to show relationships of tests or ratings. Correlations between test scores and later success in school or occupation are commonly used to predict success within known limits. In animal experimentation, correlations of performance scores with measures of deprivation or of brain injuries have proved very useful.

C. Related Items

When a group of items has been rated or ranked for some quality, such as difficulty, aesthetic quality, or emotional value, the ranks assigned by one judge may be correlated with those assigned by other judges. This procedure has been used to show linearity of relationships between items when judged by various groups, a relationship which must be known in the scaling of test items (see Chapter 23).

III. ERRORS IN APPLYING CORRELATION COEFFICIENTS

The widespread use of regression and prediction formulas makes it necessary to guard against spuriously high or low correlation coefficients. Spuriously high or low r's are sometimes obtained if regression lines and prediction formulas are used when they are not applicable to the case. Some of these errors can be avoided rather easily, but others are hard to eliminate. Correlation techniques must be carefully interpreted in order to give a clear representation of the facts.

A. Errors Which Usually Lower Correlations below Their True Value

Below are listed six types of errors which tend to make correlation coefficients lower than they would be if normal groups were carefully measured. These are:

1. GROUPING OF RESULTS IN LARGE CLASS INTERVALS. If, when making a scatter diagram, results are grouped into very few class intervals, say two or three, some scores are treated as if they were nearer the center of the group than they are. This procedure usually results in a slightly smaller correlation coefficient than would result from using fifteen class intervals, because a high correlation generally depends upon having wide and accurately determined deviations among a group of scores. A correction for this type of error is given in standard texts on statistics. The amount of correction is generally small; hence it is used for only the most careful sort of comparisons.

2. A CURVED REGRESSION LINE. It sometimes happens that regression lines are curved rather than straight, as in Figure 19, Diagrams 5 and 6. Curved regression lines may be caused by skewed distributions or by unequal steps in a scale. A product-moment correlation coefficient for such situations will be too low. The true correlation coefficient may be found by using another formula called the *eta* formula. There are several ways of finding out whether a regression line is curved or not, without actually drawing a large scatter diagram. These procedures are given in more advanced texts on statistics.

3. SAMPLES CHOSEN FROM A PART OF A GROUP. Correlation coefficients from parts of a group are usually smaller than those from the whole group. The reason for this is that in a small homogeneous group, sharp contrasts in ability are usually lacking. If all persons in a group make the same or nearly the same scores on a test, the correlation with any other test will be nearly zero. Thus, it happens that a correlation of .50 among college students or among feebleminded persons may represent about the same degree of coincidence as a correlation of .75 in an unselected population. Figure 20 shows the result of selecting a small part of a larger group for correlation purposes. Comparable results are very difficult to secure when various unrepresentative samples of population are studied.

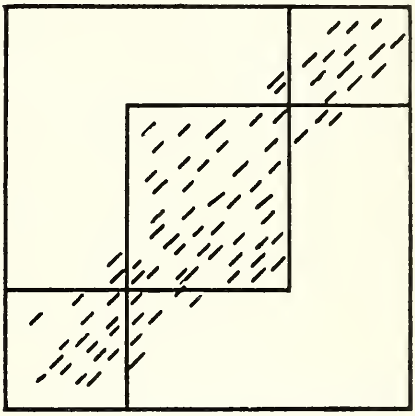


FIGURE 20

Scatter Diagram Showing the Effect of Sampling Narrow Portions of a Large Normal Group

The correlation of scores for the total is approximately .81, but the scores in the smaller squares correlate nearly zero.

4. RANDOM ERRORS. The correlation between two different measures will generally be too low because test scores or estimates are only approximations of the true scores. Random errors of measurement in one test are by definition not coincident with those in another test, and therefore, they reduce coefficients of correlation. There are several ways of correcting for this reduction which is called *attenuation,* when the self-correlations of the two tests are known (see Garrett [1937]). The corrections are small if the two tests have high self-correlations. In other words, when there are large chance variations in each of two tests, their measured relationships will show a large increase if corrected for attenuation.

5. Sampling of More Than One Process in a Test. Many tests allow the subjects to succeed by the use of different processes. Thus a high score in the Minnesota Spatial Relations test may be due to rapid manipulation, or to methodical planning or specific memory of a plan of work. This situation is usually true of all puzzles and form boards and of many verbal tests. In general, the more able persons make their scores through good planning, the less able score through rapid manipulation, and either group may use memory. The chief factor discriminating between individuals in any test will vary according to the ability of the persons taking the test. It is probable that such variations in processes lead to lower correlations between repetitions of a test than would be found if the same processes were measured in all persons in a group. A remedy for this situation is to devise tests in which success can be had by only one particular combination of processes. The construction of tests of this sort is a very difficult and challenging problem. It is discussed in many of the following chapters.

6. Use of Raw Scores Instead of Ranks. It sometimes appears that the correlation between raw score patterns on two variables is low, when the relationship between ranks as shown in an individual's profile is high. This is found, for instance, in comparing interests with vocabulary. We cannot ascertain whether a person's field of greatest interest corresponds to his field of greatest information by correlating scores, since a low score might in the case of a poorly informed person represent his field of greatest information. This problem can be solved by finding the correlation between the ranks of the test scores and the ranks of interest scores in an individual's profile. The tests may be assigned ranks by first plotting sigma scores on an individual profile, then noting their order. An illustration of the differences between correlating raw scores and correlating ranks in individual profiles was reported by Fryer (1931). The correlations between scores in academic tests and preferences for

academic subjects were very nearly zero. When the correlations between these same variables were based on ranks in individual profiles, they were much higher, with a median of .60. This was taken to indicate that a person's highest preference is likely to be his field of highest academic achievement.

B. Errors Which Raise Correlations Spuriously

There are three main types of errors which make correlations too high:

1. Sampling of Extremes. It is possible to choose persons from a group so that relationships appear higher than they would if the total population were included. This would happen if one omitted a portion of cases from the middle of the group, Figure 20, as the greatest variations between scores on two tests are usually found in the middle two-thirds of the cases. Although such selection is not common, it occurs occasionally in samples from special clinics or employment situations.

2. Sampling a Third Variable. If one compares the scores of two tests for a group composed of children who range from 6 to 12 years of age, the correlations will be spuriously high, due to the high correlation with age which nearly all skills have during periods of rapid growth. Younger children make lower scores on all tests, due to lower maturity.

A similar error is introduced when intelligence quotients (IQ's) from two tests of children of various chronological ages are correlated. IQ's are secured by dividing the mental-age score (MA), by the chronological age (CA). The division of MA's by CA's results in artificially raising the correlation, for the CA's used for each child are the same in both tests.

There are two methods of avoiding these errors. The common procedure is to correlate only the scores from separate age groups. The other method is to calculate what the correlation would be if the effects of age were removed. This can be done by the partial correlation technique shown in any good statistical manual. Partial correlations are not advised for use if the spurious factor can be removed directly.

A similar situation is often found in interpreting a correlation between any two tests. Thus, speed-of-reading tests and arithmetic tests both involve co-operation with the examiner. If poor co-operation and poor ability go together, as is often the case, the correlations found between scores on the two tests will be higher than pure measures of ability would be. This problem of analyzing the factors which raise or lower correlations is a very complex and persistent one. It is discussed in considerably greater detail in Chapter 7.

3. Correlating a Part with the Whole. In many situations one wishes to find out whether a part of a test is consistent with the whole test. The correlations between single items and total scores are likely to be spuriously high if the item makes up a considerable part of the total. Corrections are not large enough to be considered essential if the item to be evaluated is one of 10 or more items each of which has equal weight in the total.

IV. HOW LARGE MUST CORRELATIONS BE FOR THE PURPOSE OF PREDICTIONS?

A very practical question is often asked, "Is this correlation high enough to be useful for predicting individual successes or failures?" From the discussion above it is clear that no definite limits can be set, and that a correlation must be interpreted in its total situation. One must always take into account these factors which have just been discussed because they may introduce errors into the correlation coefficients. If errors have been eliminated as far as possible, then there are two ways of evaluating the significance of a correlation. One shows the probable error of correlation; the other gives an indication of how much the standard-error-of-estimate is reduced by an increase in the size of the correlation.

A. Probable Error of a Correlation Coefficient

The probable error of a correlation (PE_r) is used to indicate the extent of random sampling errors. It shows how much the coefficient will probably vary on a chance basis, that is, if the same or similar groups of persons were measured in the same way many times. The PE_r shows the theoretical range of the middle 50 per cent of r's obtained by random sampling. If an r is four times its own PE, it is highly improbable that the r's true value will be as low as zero, since this would occur less than once in a thousand trials. This is clear, when one considers the dimensions of a normal distribution curve, Figure 11, which shows that .99904 per cent of the total area will be included on a base line which is 8 PE long, that is, 4 PE each side of the mean. Both an increase in the total number of cases in a group and an increase in the size of the correlation would result in a decrease in the size of the PE. The calculation of the PE_r is shown in Table 14. The PE_r ranges from .0021 when N is a thousand and r is .95, to .1508 when N is 20 and r is .00, as shown in Table 16. From this table it appears that the PE for the correlation shown in Table 14 is approximately .017.

The PE_r tells the probability that an r will vary. Thus, if two coefficients, one of .50 and the other of .90, have the same PE, they are equally variable measures. This is a valuable bit of information since it tells how much the correlation coefficient may be expected to vary upon retrials. It does not, however, tell how much better the prediction for individual cases is when the r is .90 than when it is .50. The inaccuracy of a prediction is indicated by the coefficient of alienation.

TABLE 16

Probable Errors of the Coefficient or Correlation for Various Numbers of Measures (N) and for Various Values of r

Number of Measures	*Correlation Coefficient r*						
	0.0	0.1	0.2	0.3	0.4	0.5	0.6
20	1508	1493	1448	1373	1267	1131	0965
30	1231	1219	1182	1121	1035	0924	0788
40	1067	1056	1024	0971	0896	0800	0683
50	0954	0944	0915	0868	0801	0715	0610
70	0806	0798	0774	0734	0677	0605	0516
100	0674	0668	0648	0614	0567	0506	0432
150	0551	0546	0529	0501	0463	0413	0352
200	0477	0472	0458	0434	0401	0358	0305
250	0426	0421	0409	0387	0358	0319	0272
300	0389	0386	0374	0354	0327	0292	0249
400	0337	0334	0324	0307	0283	0253	0216
500	0302	0299	0290	0274	0253	0226	0193
1000	0213	0211	0205	0194	0179	0160	0137

Number of Measures	0.65	0.7	0.75	0.8	0.85	0.9	0.95
20	0871	0769	0660	0543	0419	0287	0147
30	0711	0628	0539	0444	0342	0234	0120
40	0616	0544	0467	0384	0296	0203	0104
50	0551	0486	0417	0343	0265	0181	0093
70	0466	0411	0353	0290	0224	0153	0079
100	0391	0345	0294	0242	0187	0128	0066
150	0318	0281	0241	0198	0153	0105	0054
200	0275	0243	0209	0172	0133	0091	0047
250	0246	0218	0187	0154	0118	0081	0042
300	0225	0199	0170	0140	0108	0074	0038
400	0195	0172	0148	0122	0094	0064	0033
500	0174	0154	0132	0109	0084	0057	0029
1000	0123	0109	0093	0077	0059	0041	0021

(Garrett, 1937. By permission of Longmans, Green and Co.)

B. Coefficient of Alienation (k)

The prediction of the most probable score in one variable from a score in another, it will be recalled, was made by means of the regression equation. The variability of the most probable score was then evaluated by means of the standard-error-of-estimate. Since the standard-error-of-estimate was calculated from the correlation coefficient, it is possible to show the relationship between these two indices. If the correlation between two variables, X and Y, is zero, then the standard-error-of-estimate ($\sigma_{Y \cdot X}$) is the same as the standard deviation for the whole Y distribution. In this case, the errors in predicting scores on one variable, Y, from those on another variable, X, are maximum. The only prediction that can be made is that one's Y score will be somewhere among all the Y scores, and that one's most probable Y score will be the mean score. If the correlation is larger than zero, the standard-error-of-estimate will be smaller than the standard deviation of Y. The proportion by which the standard deviation of Y is reduced is called the coefficient of alienation. This coefficient is equal to $\sqrt{1 - r^2}$ and its magnitude for various correlation coefficients is shown in Table 17. Reading from this table, the coefficient of alienation for the correlation coefficient of .451 in Table 14 is approximately .898. In other words, the standard-error-of-estimating a child's height is .898 of the standard deviation of the entire group of children.

TABLE 17

Coefficient of Alienation *k* for Values of *r* from .00 to 1.00

r	*k*
.00	1.0000
.10	.9950
.20	.9798
.30	.9539
.40	.9165
.50	.8660
.60	.8000
.70	.7141
.80	.6000
.866	.5000
.90	.4539
.95	.3122
.98	.1990
.99	.1411
1.00	.0000

(Garrett, 1937. By permission of Longmans, Green and Co.)

It also appears from Table 17 that the correlation must be .90 to reduce standard-error-of-estimate to 45 per cent of the standard deviation of the total distribution. The correlation must be .995 to reduce it to 10 per cent. This means that it is wise to be cautious in placing faith in individual predictions made when the correlations are less than .90. Since most of the correlations reported between raw scores and various criteria of success are in the neighborhood of .50, predictions for individual success on the basis of a single comparison are usually too far from accuracy to be very useful. This fact does not discourage the careful worker, but makes him realize the need for very much more precise measures.

V. THEORETICAL INTERPRETATIONS OF CORRELATIONS

In addition to these practical indications of the significance of correlations for predictions, the fact that a correlation coefficient is merely an indication of a coincidence in time and place must be remembered. A correlation simply shows the relative positions of items or persons in specific groups; it never shows why they appear in these positions. Any interpretation of cause or effect between correlated variables is entirely independent of the correlation itself, and must be supported on the basis of first-hand observations of behavior. This fact is often forgotten in the shuffle of figures, and as a result a correlation is erroneously supposed to indicate the extent of causal relationship. It will only do so when the relationship has been demonstrated by a well-controlled experiment.

This discussion leads to the question, "What sorts of relationships may be shown by a high correlation?" A good many answers have been made, depending upon various assumptions concerning the nature of the behavior variables and their combinations. There are two main types of interpretations of correlations. One interpretation assumes that a high correlation is caused by the presence of identical elements in the two measures; the other, that it is caused by the covariations of nonidentical elements.

A. Identical Elements

A high correlation may be interpreted as due to the presence of identical elements or factors in the two variables only when such elements have been demonstrated. Thus, an arithmetic test and an algebra test might be thought to correlate highly through the presence of identical number concepts. This may be true, but several have pointed out that it may be false, and that the high correlation may be due to entirely different processes.

A general case is illustrated in Figure 21 in which it is assumed that the scores on X are determined by amounts of two independent elements N and P, and the scores on Y by amounts of N and R. If it is further assumed that the score is simply a sum of the component elements rather than some more complex combination, then the correlation between X and Y will be (Garrett, 1937, p. 348)

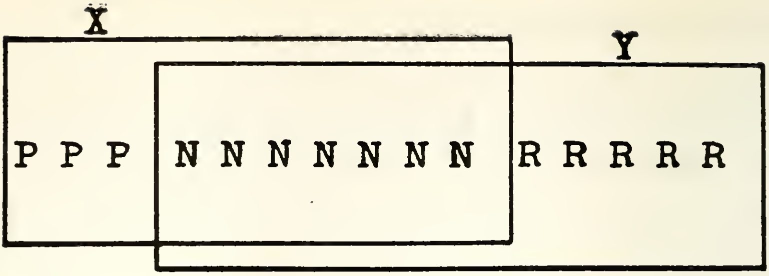
$$r_{XY} = \frac{N}{\sqrt{N + P}\,\sqrt{N + R}}$$

when N, R, and P are the amounts of various elements.

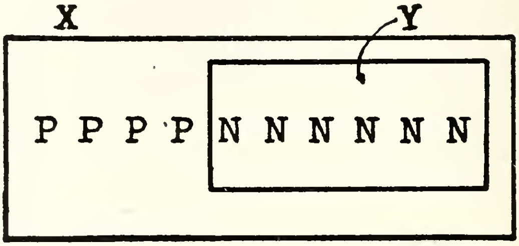
The simplest case of this sort is given when X includes all of Y, Figure 22. Here, the square of r always gives the proportion of the elements of X which determine the score of Y, and r^2 is called the *percentage of determination.* This interpretation, that the square of a correlation coefficient is a percentage of determination, is not one which is usually justifiable, for the factors which determine the scores under consideration are seldom known. There is no method for finding the percentage of determination when the factors are unknown.

B. Covariation

The second interpretation, which is sometimes more applicable to test results, is that a high correlation is due to a coincidence of two or more nonidentical variables. This is called *covariation.* The scores in X may be highly related to the scores in Y when the persons who have ability to succeed in X also have ability to succeed in Y, Figure 22. For instance, knowledge of physics is generally found to be un-

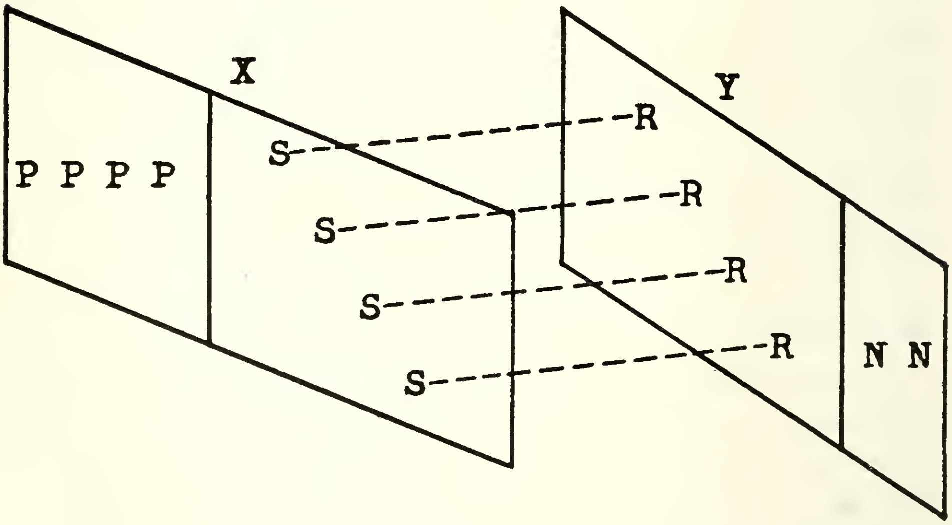

X
Y
P P P
N N N N N N N
R R R R R

A. $$r_{xy} = \frac{N}{\sqrt{N+P}\quad\sqrt{N+R}} = .64$$


X
Y
P P P P
N N N N N N

B. $$r_{xy} = \frac{N}{\sqrt{N\,(N+P)}} = .77$$

FIGURE 21. **Correlations When Identical Independent Elements Are Coincident**


X
Y
P P P P
S
S
S
S
R
R
R
R
N N

C.

FIGURE 22. **Correlations When Nonidentical Independent Elements Correspond**

related to singing, but if all persons in a group had covariant training and interests in physics and singing, then the correlation between tests of these two abilities might be high. In such a case, it would appear that the abilities were coincident, but it would not be safe to infer that they were based on identical processes. Two explanations are commonly made. Some persons are inclined to say that there is a general innate ability to learn which can be directed into either field, though this assumption is certainly not justified on the basis of the correlation results alone. Others would favor the explanation that the two test scores are determined by a complex combination of an unknown number of factors, many of them cultural. These hypotheses are discussed in Chapter 7.

5

MEASURING INSTRUMENTS

The previous chapters have laid foundations for a discussion of measuring instruments. These foundations include knowledge of types of appraisals of behavior and familiarity with techniques for comparing the scores of various persons. This chapter defines the main types of instruments and describes several characteristics which are considered to be assets in any test or estimate. Methods by which desirable characteristics may be secured are outlined here, and illustrated more completely later in the book.

I. TYPES OF TESTS

A. Definitions

The words *test, measuring instrument,* and *scale* are used somewhat interchangeably, but they have distinct meanings of their own. The word *test* may refer either to an examining procedure or to the materials which are used in an examination. The materials may be printed forms, objects, sounds, or other stimuli to which one must respond. These are selected to aid the examiner in discriminating between various abilities and degrees of ability. A *test item* is a prescribed stimulus which when used in an examination yields a unit score. A *measuring instrument* is a device which can be used to evaluate a type of structure or performance with a standard set of values, called a *scale*. A scale is a numerical scheme of reference consisting of points or steps which are usually equivalent in some respect.

B. Formal and Informal Tests

There are three main differences between an informal test and a formal or standardized test: (1) The formal test has more rigidly standardized directions for administration and scoring which come from the revision of many preliminary trials. (2) The context of a formal test has been more thoroughly scrutinized to include only important facts and skills and to eliminate ambiguities and chance factors. (3) The formal test has usually been widely applied so that norms for many persons in various age, grade, or occupational groups are available.

These differences are all in favor of the formal test, but an informal test is often preferred because it can be constructed to fit the needs of a particular situation more aptly than any formal test available. This situation is so frequently the case in industry and in progressive schools that employment specialists and teachers should be alert to construct their own tests on many occasions. All of the material in this chapter and the next is designed to help one construct and evaluate either formal or informal tests. Mental tests may be classified into three main groups according to their uses:

1. To measure present ability; achievement tests
2. To predict success; aptitude tests
3. To diagnose behavior; tests of psychological processes

C. Achievement Tests

Achievement tests are designed to measure skills and information which have been learned, either in particular courses of training or from experience elsewhere. There are widely used tests for nearly all school subjects at various grade levels (see Chapter 9). Test items for achievement tests are generally selected from those which have been used by teachers as indications of success in a course of instruction. If a large group of competent teachers agree that the data for a test item have been taught in a course on American history, then that item is considered relevant for a test in this subject. The first criterion of exclusion or inclusion is the agreement of competent judges on its relevancy.

A group of progressive educators has set up ideals which are much more inclusive than the mastery of certain subject matter. A typical list of these ideals is given by Raths (1938, p. 90) who summarized the work of the evaluation staff of the Commission on the Relation of School and College, of the Progressive Education Association, as follows:

1. Improved habits and abilities with relation to reflective thinking.
2. Wider and richer interests.
3. An increasing consistency in important attitudes.
4. Increasing facilities to adjust socially.
5. Developing creativeness.
6. Improved study skills and work habits.
7. An increasing stock of vital information.
8. A wider and better appreciation of literature, the arts, and music.
9. A developing sensitivity to socially significant problems.
10. A functional philosophy of life.

The construction of precise instruments to measure progress toward all of these objectives is a gigantic task, but operational definitions have been provided for most of these ideals and marked progress is recorded in various chapters in Part II.

D. Aptitude Tests

A fairly large number of tests have been called aptitude tests by their authors. Tests designed to indicate clerical, mechanical, artistic, musical, and language aptitudes are described in Part II. Inspection of these shows that they demand much the same sort of skills and information as achievement and psychological tests. Often the same items are found in all three types. Although there is no agreement among psychologists on one technical meaning for the word *aptitude,* there are three fairly common meanings:

1. Aptitude tests are any tests which happen to predict later success to some degree. It must be demonstrated that high standing on the test will indicate great success after some specialized training. This point of view is voiced by Bingham (1937) and by other authors.
2. Aptitude tests are those measures of achievement which will predict some future development, but which will not themselves increase or change significantly with any further development. This definition is used by Paterson and Darley (1936), who offer the Minnesota Clerical Test as an illustration, Figure 2. This type of definition generally implies that all persons tested have, through previous maturation and experience, reached fairly high levels of proficiency in the skills tested. Hence, those with high scores have demonstrated their ability to develop in work of a similar nature.
3. A third type of aptitude indicator is not a single score, but a series of scores arranged in a curve of individual development. This indicator is seldom used, for it is difficult to secure records for a number of persons which are comparable in amounts of practice and motivation. For the most careful predictions, developmental curves would certainly be much better than one test score. Because of this, there is a strong emphasis in mental measurements today upon securing curves of long-time development.

When a test is to be used to predict success in an occupational or professional school, it should, according to one theory, include items which measure all the factors which contribute to success and give these items about the same weight in the test score that they have in producing success. However, it has been found to be very difficult to decide what factors are important in the development of occupational successes, since the tests are administered at a period when persons have had very little occupational training. In order to construct an aptitude test which will give good prediction, a program must be followed which will:

1. Select a fairly large variety of items which seem to have some predictive value. The selection of tests is usually aided by a careful analysis of the skills of those experienced in an occupation
2. Apply these tests to a large number of persons who are just beginning training for an occupation or profession
3. At yearly periods after training has been finished, secure ratings or other indications of occupational success
4. Compare the ratings of success with the various test scores
5. Combine the test scores to give the best prediction of success. Often a number of the original tests are eliminated before a final combination of tests for an aptitude scale is made

This procedure has been followed with very good results in a few industrial and business situations.

E. Psychological Tests

Psychological tests are primarily designed to demonstrate the truth of a hypothesis concerning some aspect of mental ability. Items for these tests are selected according to the concept of a particular ability which an author wishes to investigate. This selection results in a preliminary form which is scrutinized for chance factors and ambiguities, as described in the next chapter, and also for evidences that the test items are actually measuring what they were designed to measure. Tests of this sort are described in Chapters 10, 11, and 12.

F. Altitude, Speed, and Breadth Tests

The three types of tests just described may be reclassified according to altitude, speed, and breadth. *Altitude* tests, which are sometimes called *power* tests, allow a liberal amount of time so that all persons tested will attempt nearly all the items which they could possibly pass. The scores of an altitude test yield an indicator of the highest level which a person achieves. The best altitude tests-are constructed from carefully scaled items which increase in complexity as the test progresses.

Speed tests, which are also called *rate* tests, are of two sorts. One, called a time-limit test, is composed of items which are all of similar difficulty, as in the name comparison test, Figure 2, page 21. A time limit is used which is so short that none of the group can finish the test. The second, called a work-limit test, requires a standard series of operations, and the score is the time needed to complete the operations at a given level of excellence. This is illustrated by an assembly or a form-board test where materials are to be put together to form a pattern.

Breadth tests are designed to measure one's range or variety of skills or information. Great speed is not required, and complexity is fairly constant. A good illustration of this type is the Co-operative Test of Current Events, which samples a person's knowledge of recent developments in five fields: Science, Economics, Foreign Affairs, Arts, and Recreation.

Many tests are mixtures of altitude, rate, and breadth items; but mixed tests are not desirable, for they do not allow clear interpretation of results. For instance, when an altitude test is given with rather short time limits, two persons may get the same score for different reasons. One may be a fast yet superficial worker who makes his maximum score during the time allowed. Another person may have greater ability, but work so cautiously that he is just getting nicely warmed up when time is called. By a similar logic, items which are carefully graded for complexity of thought for an altitude test should not be mixed with those which are simply rare, although rare items may be appropriate in a breadth test. Likewise, the breadth items should not be given with short time limits since the main point of a breadth test is to enumerate the various facts which a person knows. Frequently one finds interesting temperament or motivation indices in the relative speed at which persons work, so that a record of this factor in any sort of test is often desirable.

The usual method of establishing time limits involves trying out items upon typical groups of persons. This is done in all cases of well-standardized tests. Altitude tests have been found to give adequate samples of ability if enough time is given to allow approximately 90 per cent of the group to attempt every item. This is particularly true when the tests depend to a large degree upon information; when the items have been arranged in order of difficulty; and when the range of difficulty is such that at least 5 per cent of the group can succeed on the hardest items, and 5 per cent will fail on the easiest items.

Two reports on the same test are interesting in this connection. Yerkes et al. (1921) found that 510 soldiers in the United States Army had nearly the same relative position in two Army Alpha tests, one of which was given with standard time limits, and the other with double the usual time. The correlation was .965. Greene (1927) repeated this experiment with 216 college students and found a correlation of only .541. In another group of 110 college students the correlation was .14 between two Army Alpha tests, one of which was allowed the standard time, and the other, unlimited time. Some students took four times as long as others. In any test, speed factors become more important in discrimination of individual abilities when complexity is reduced or when more able groups are tested.

II. CHARACTERISTICS OF A GOOD INSTRUMENT

Although the various types of tests have different uses and forms, there are four characteristics that are considered to be assets in all measurement: validity, reliability, equivalence of steps, and practical aspects.

A. Validity

The word *validity* has two somewhat different meanings, one qualitative and the other quantitative. If one chooses the qualitative meaning, then in order to be valid a test should measure only what it is supposed to measure. A good test of temperature should not be affected by pressure or oxidation, and a good measure of color vision should not vary with degree of reading ability. To assure the validity of a test, one must include or sample all relevant items and exclude all irrelevant items. In some cases it is fairly easy for competent judges to agree upon relevant and irrelevant factors, but in many cases it is difficult. In all cases the final decision must come from some procedure which defines relevance or validity in terms of operations which have been agreed upon. Thus, a valid test of intelligence is one which distinguishes correctly between those who are judged by competent judges to have various degrees of this ability. It often happens that a refined measuring instrument seems to indicate an ability more accurately than the most competent judges. However, if both are appraising the same ability exclusively, both appraisals are qualitatively valid; they simply vary in random errors. *Random errors* are operationally defined as those which change individual scores without changing group norms on two trials of the same test. In large samples random errors are normally distributed.

1. Objectivity. An indication of the qualitative validity of an appraisal is its objectivity. The term *objectivity* has acquired a special meaning when applied to measures or ratings. A highly objective appraisal is one which is not

affected by the special idiosyncrasies of the appraiser. The opposite of objectivity is subjectivity. Hence any test is free from subjective variations if the judges involved agree in their judgments and in the credits assigned to each performance. Large subjective variations are often found in appraisals of artistic ability, social co-operation, and in essay types of mental tests, for in these cases standards of excellence are difficult to agree upon.

Short-item tests of the true-false, multiple-choice, or completion type are sometimes called objective type tests. This usage is misleading since the type or form of item does not guarantee that it represents universal or widespread agreement of competent judges. A well-standardized test is usually composed of items upon which large numbers of competent judges agree, but many short-item tests contain very controversial items.

Adams (1936) described a method of appraising the subjectivity of a series of ratings. Ratings were considered to be objective if two ratings by one person correlated as well as and no better than two ratings by two different persons. For instance, if ten judges each rated fifty paintings for artistic composition twice, subjective variations would be indicated when the two ratings of a single judge correlated .90, and the mean correlation between ratings of any two judges correlated .70. If, however, both of these figures were the same, then there would be evidence that all judges had probably used the same or similar standards, and the ratings would be considered to be objective. This method of appraising subjectivity should be used more frequently, for it is easy to apply, both to ratings and to tests which are supposed to be measuring the same patterns. Since nearly all intelligence tests show higher self-correlations than correlations with other tests of intelligence, one must conclude that there is a subjective factor in the composition of such appraisals. Adams' method of appraising objectivity is not conclusive but permissive. It can be used to show that subjective differences were present, but not to prove that all judges have used the same standards.

If one chooses the quantitative concept, a test is considered to be valid when it is a very accurate measure of that which it purports to measure. This definition is the one usually encountered in published test results. The commonest check on validity of a test is a correlation between the test and some other criterion of the behavior which the test is designed to measure. If this correlation, called a *validity coefficient,* is approximately 1.00, the test is thought to be highly valid in either the qualitative or the quantitative sense; if low, to be invalid. Validity coefficients are widely furnished from standardized tests, but some precautions are necessary for correct interpretation. Thus, a high correlation is no conclusive proof of validity, for a test may be measuring some pattern of behavior which is not the same as its criterion, yet which happens to be highly correlated with it in the group tested. Another difficulty in interpretation lies in the fact that a low correlation between a test and its criterion may be found when both are appraising the same pattern, but when one or both have large random errors. Hence a low validity correlation does not necessarily mean that a test is qualitatively invalid. A low correlation means that the prediction of one variable from the other is poor.

2. SAMPLING METHODS. A check upon the validity of a test may be applied when one wishes to find out how well two small samples represent a large sample of test items. This check is important since mental tests must from necessity usually be short. Modern information tests are often made by selecting about 40 items from among several hundred. To check the validity of two short tests which have been made of items chosen from a large number, one must apply them both to a fairly large group of persons who are thought to possess wide differences in the information tested. Next, the scores from the two forms are correlated. When

correlations are low, one must conclude that some factors influence the scores on one test, but do not influence the scores on the other test in the same way. A factor is defined as any aspect of behavior or of the situation. This correlation procedure is necessary to eliminate poor samples, but it still does not guarantee that good samples of the master list have been secured, for factors within the master list may correlate highly with other factors not included in it. Such extraneous items or factors may creep in during the construction or administration of various test forms. It is further necessary to prove that nothing but the factor one desires to measure is represented by each test score. It is impossible to prove this with a simple correlation technique, for a correlation coefficient never shows mental processes but simply coincidence in relative position. For more adequate proof, one must use direct observation to appraise the behavior represented by the test scores.

B. Reliability

The discussion of validity has brought out the fact that a correlation between two tests may be low owing either to the difference in the patterns they are measuring or to the presence of random errors of measurement. If it were possible to repeat a test upon the same group under conditions which allowed only random variation, the correlation between two trials of the same test would show the extent of random variations. Such conditions are very rare, but many workers have attempted to find out how much random variation one might expect in applying a test to a particular group a number of times. These attempts yield what are usually called *reliability coefficients.* Three coefficients are commonly found: retest reliability, split-half reliability, and equivalent-form reliability coefficients. Since these procedures often give different results, they should not be confused. In all cases, a high coefficient is considered to be an indication of small random variations, and a low correlation indicates either random or systematic[1] variations or both.

Retest reliability is secured by giving a group the same test twice within a few days. When it is difficult to secure two tests of the same group, a split-half reliability may be secured by correlating scores on one half of a test with scores on the other half. Often the scores for odd-numbered items are correlated with the scores for even-numbered items. Correlations from the halves of a test, given on the same day, are frequently from .05 to .10 points higher than correlations from the same halves given a few days apart. The differences are due to changes in attitudes or conditions of the persons tested.

The correlation between halves of a test indicates only the variations between each half. Since the test is to be used as a whole, it is desirable to know the probable reliability of the test as a whole. This can be calculated from the Brown-Spearman (Spearman, 1910) formula for prophesying the probable effect of lengthening a test. The self-correlation of two tests which are made n times as long as they were originally, where r_{11} is the known correlation of the two originals, would be (Garrett, 1937, p. 315):

$$r_{nn} = \frac{nr_{11}}{1 + (n - 1)r_{11}}$$

For example, if the two halves of a test correlate .60 with each other, the whole test would probably correlate .75 with another test of similar length and design. That is:

$$r_{nn} = \frac{2 \times .60}{1 + (1 \times .60)} = \frac{1.20}{1.60} = .75$$

This formula may be used to find how long a test must be made to achieve a particular self-correlation. Of course, the formula would not apply if, by lengthening a test, one introduced new factors, such as fatigue, loss of interest, new sorts of items, or practice on the part of some individuals.

---

[1] Systematic variations are defined as those which change the group norms as well as individual scores on two trials of a test.

The equivalent-form reliability coefficient is secured by correlating two forms of a test that are supposed to be sampling the same ability. Equivalent-form reliabilities are usually .05 to .10 points smaller than retest reliabilities, for equivalent forms have different specific items, while a retest uses the same items as the original test. Retest reliabilities are also sometimes inflated by specific memories of previous performances; hence equivalent-form reliabilities are preferred for test appraisal.

INTERNAL CONSISTENCY. One characteristic of a test which is often considered as evidence of reliability is called *internal consistency*. High internal consistency is defined operationally by a situation in which each item or subtest of a battery is found to arrange a tested group of persons in the same order of excellence as that secured by the total score. This characteristic is always a valuable one for tests which are constructed to appraise one independent factor. It may not be appropriate for aptitude tests of complex sorts, such as sales aptitude or engineering aptitude, where success may be thought of as the combination of several unrelated factors. Of the various methods for evaluating internal consistency, two that are widely used will be described. One is called the *split-group* method, and the other the *correlation* method.

The split-group method first divides a large group of persons into thirds, or quarters, on the basis of their total scores on a test of many items. Then the per cents of persons in the highest and lowest thirds who pass each item are found. If the lowest third does as well or better on an item than the highest third, one must infer that the item and the total test do not measure the same processes. If all of the highest third pass the item, and all of the lowest third fail it, then the item divides the group into classes much as the whole test divides the group.

This split-group method is illustrated by Anderson's (1935) report of an analysis of a 222-item examination in Educational Psychology. On the basis of discrimination between lower, middle, and upper thirds, 86 items were classified as good, 83 as poor, and 53 as intermediate. The scores of each student on good items and poor items were calculated. The means of the good and the poor items were nearly the same, 50.3 and 55.3. But the standard deviations of the good items were 11.1, and of the poor 3.5; and the scores on the total 222 items correlated with the scores on the good items .95, and with the 83 poor items .45. The correlations of final grade in the course with the good items was .90, and with the poor .22, and with the total score .89. The split-half reliability for the total examination was .864, for the good items .832, and for the poor items .06.

These figures show that the split-group evaluation allowed a selection of approximately one-third of the total items, which gave as good discrimination as the whole examination, and which predicted final grades in the course a trifle better than did the whole examination. It was further shown that the total items do not correlate quite so well with high-school rank and scores on the Iowa English Examination as do the good items alone. The study might well be followed up by work to show the smallest number of good items which would be needed to allow good discrimination of this group and of other similar groups.

Another illustration of the use of this method is that of the work of Terman, *et al.* (1917), Table 55 A, page 279. Here the per cents of high, low, and medium IQ groups who passed each item are indicated. Since the CA's were nearly constant for each group, the IQ in this case represents the total score of the test. Nearly all the items show large differences between the lowest group and the middle group. The differences between the middle and the highest group are usually smaller owing to the fact that the items were chosen so that the middle group would have more than 50 per cent correct. Items which did not show as large discriminations were omitted from the 1916 Stanford Revision of the Binet Test.

A test may also be constructed to have high internal consistency by comparing each item with some criterion of success other than the total test score. Such a test may have the added advantage of being a valid predictor of some practical success. For instance, Uhrbrock and Richardson (1933) divided a group of supervisors into thirds on the basis of estimates of efficiency. The proportion of persons in each third who passed each of 820 items was found. The 85 items which best discriminated among the thirds were found to correlate .71 with supervisory ability, whereas a test of 400 unselected items predicted supervisory success with a correlation of only .49. This study shows that by this method the best items for predicting success can be picked out rather quickly.

The correlation method, which is preferred since it seems to involve less work, correlates the scores made on each item with the scores from the whole test. If such a correlation is high and positive, it is likely that the item discriminates between persons in the same manner as the whole test. If the correlation is low or negative, success on the item is likely to be determined either by chance or by other factors than those which affect the test as a whole.

The correlation method was followed by Greene (1939) in devising vocabulary tests for adults in eight different fields. The items which did not correlate with the total test score by at least .30 were discarded. Mean correlations were approximately .40. Thus each item used was known to predict the total score much better than chance. This amount of internal consistency is unusually high for a test of this sort confined to college groups.

Vincent (1924), using college entrants, found that the separate correlations between 365 single items (completion, arithmetic, and comprehension, from Thorndike's CAVD test) and a two-hour intelligence test ranged from .00 to .82, mean .32. For another group of 70 items with sixth-grade pupils, similar figures ranged from approximately .15 to .80, mean .33. Thorndike (1926) also reported correlations between reading and vocabulary items with a long intelligence test. The figures range from −.45 to .80, median .38. These items were apparently not used by Thorndike in the selection of items for the CAVD tests, but they were reported to show the facts.

The inclusion in a test of items which have low correlations with any criterion produces a measuring instrument which has low internal consistency, and prevents any careful analysis of mental processes. One cannot safely conclude from such tests that there is any general pattern of behavior or skill present; the opposite is more likely.

This discussion of test validity and reliability is merely introductory. It points to the need for better methods of appraising the worth of a test which will clearly distinguish between qualitative and quantitative aspects. No simple method is likely to appear but interested students will find more extensive practical and theoretical discussions well summarized by Brown and Thomson (1925), Thurstone (1935), Adams (1936), Guilford (1936), and Garrett (1937).

C. Equivalence of Steps

Another characteristic of a good measuring instrument is the equivalence of units in the scale which is used for recording the results. Obviously, if a yardstick were made up of inches some of which were three or four times as long as others, the results of measuring the length of a person would vary, depending upon the part of the yardstick which was used. Similarly, if a test of ability has some items which are much harder than others, its use is likely to result in errors of measurement. Figure 23 shows three possible arrangements of units which are probably common in mental tests. Arrangement A is typical of all tests in which the items become progressively more difficult. Arrangement B is typical of spiral-omnibus tests which consist of items from different fields mixed together (see Table 63, page 322). Arrangement C is typical of some speed tests, in which a task such as digit-symbol association, page 314, may become easier as one progresses.

The problem of establishing equivalence is a perplexing one; there are several varieties of equivalence and several methods for deriving scales with equivalent units. Chapter 10 describes scales used when equivalence is defined as equal periods of growth. Chapters 11, 22, and 23 consider scales in which equally-often-noticed differences are considered equivalent. Although these two methods of scaling do not yield similar results, both are useful.

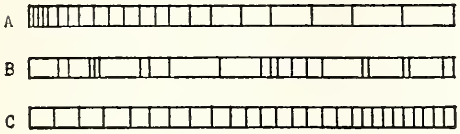


FIGURE 23

Unequivalent Scale Items

Along with the need of equivalent steps in a scale, one also needs equivalent zero points for the various scales. Some workers favor the use of a temporal zero which places the beginning of mental development at or near the time of birth. Others have postulated a zero of an ability as a point on the scale at which no variation of that ability is found in a large group of measures. This concept is discussed in Chapter 22.

D. Practical Aspects

For practical purposes a good testing instrument should require minimum time, cost, and effort in its administration, scoring, and interpretation. Obviously, as the testing situation varies, the relative importance of these aspects will change. A good testing instrument should also allow a good sampling of the abilities of a person without disturbing or embarrassing him. In order to avoid coaching and to provide for growth studies, two or more equivalent forms should be available. A list of practical requirements would include the following:

1. Administration
    - *a.* Optimum difficulty of items
    - *b.* Short time allowance
    - *c.* Little or no supervision, except for recorded observations
    - *d.* Simple clear directions
    - *e.* Few and inexpensive materials
    - *f.* Several equivalent forms
    - *g.* Situations of interest to the examinee
2. Scoring and Interpretation
    - *a.* Little judgment required of the scorer, machine scoring preferable
    - *b.* Only the simplest calculations required in obtaining scores, and easy methods of checking tabulations
    - *c.* Results which compare a person with those of his own age, sex, and status, grouped in terms of an index of group dispersion
    - *d.* Evidence of equivalent steps in scaling
    - *e.* Evidence of true zero points

Some of these requirements conflict with other requirements. For instance, great economy of administration time usually reduces the significance of the results. Since these practical aspects are dealt with in many places in Part II of this book, only the first aspect will be discussed here.

THE OPTIMUM DIFFICULTY OF ITEMS. Difficulty of an item is usually defined for a particular group by the percentage of persons in the group who succeed on that item. A single level of difficulty is defined as a group of items, all of which are passed by about the same proportion of a particular group. In test construction the question of what levels of difficulty shall be included always arises. If one desires to place each individual on a scale with great accuracy, then items from all levels of difficulty must be included in the test in large enough numbers to avoid chance effects. This is illustrated by nearly all the well-standardized scales which have wide ranges of difficulty. On the other hand, if one desires merely to divide a group of persons into a small number of divisions on the basis of ability, then only those items are needed which will indicate the division boundaries. The simplest situation would be one in which a group of two persons is to be divided into an upper and a lower half. The optimum difficulty would be represented by those items which one person passed and the other failed, for items which were failed or passed by both persons do not serve to distinguish between them. A more complex situation would be one in which a large group is to be separated into an upper and a lower half. In this case the only items needed would be those which most clearly discriminated between the persons in the upper and lower halves. Such items would be neither the easiest, for they distinguish only between the least able persons, nor the hardest, for they discriminate only among the most able. The items which are passed by 50 per cent of the group would most adequately separate the group into halves.

If one wishes to divide a group into quarters, then theoretically he should include in his tests, items which are passed by 25 per cent, 50 per cent, and 75 per cent of the group. When test items have been applied to single age groups in particular environments, scales which will have particular degrees of difficulty may be constructed with considerable accuracy.

6

CONSTRUCTION AND EVALUATION OF TEST ITEMS

I. TYPES OF TEST ITEMS

Items may be classified into various types according to the number and arrangement of their elements. The type of an item is important since this determines to a large extent the economy of administration and scoring, and the presence of factors causing chance errors. To a smaller extent, the type of question also determines the processes needed for success.

It should be stated before proceeding with this discussion that, regardless of the type of question, the best test for a particular appraisal is that which has the largest proportion of valid items in it. Validity, discussed in Chapter 5, cannot be known very definitely before the test has been tried out on a fairly large sample of persons. Suppose, however, that a number of item types which cover the same material have been tried out on a group and that the various types all show significant correlations with the criteria. It is then reasonable to ask which type of test question is the best to use in a particular situation. Four types of commonly found items are illustrated in Table 18.

TABLE 18

Types of Test Items

1. True-False Items. These require a judgment that a given statement is true or false as:
The sum of 6 and 7 is 13. *T F*
2. Completion-Type Items. These direct the subject to complete a picture or a statement by supplying an appropriate element as:
The sum of 6 and 7 is ______
3. Multiple-Choice Items. These call for a choice of a correct answer to a question from several incorrect answers as:
The sum of 6 and 7 is 12 15 13 17 14.
4. Matching Items. These demand that each of a list of elements, usually about 10, be matched for significant relationships with elements chosen from another list as:
Write the number of the book in front of its author's name:

Author	Book
______ Poe	1. Faerie Queene
______ DeFoe	2. Plain Tales from the Hills
______ Dickens	3. Pickwick Papers
______ Spenser	4. The Gold Bug
______ Tennyson	5. The Taming of the Shrew
______ Byron	6. Don Juan
______ Kavanagh	7. The Four Hundred
______ Kipling	8. The Idylls of the King
______ Conrad	9. Lord Jim
______ Shakespeare	10. The Odyssey
	11. Robinson Crusoe
	12. The Pearl Fountain

II. MERITS OF VARIOUS TYPES

If the purposes and time limits of a testing program are definitely fixed, then answers can be given to the question, "Which is the best type of item?" Some of the answers are summarized by ratings in Table 19. These ratings represent this writer's opinion, based on numerous observations as well as on a number of studies by other investigators. Such ratings depend to a marked degree upon personal experience; hence they might vary considerably if made by another judge. Thus, the ease of composition of a particular type of item is dependent, at least in part, on one's experience. To a lesser degree, judgments of ease of administration and

TABLE 19

Merits of Types of Items

(1 is the highest rank; 2, next highest; and 3, lowest)

	Type of Item					
	1 *True-False*	2 *Completion*	3 *Multiple-Choice*	4 *Matching*	5 *Rearrange*	6 *Essay*
1. Easy to compose	2	2	3	2	2	1
2. Easy to understand directions	1	1	1	2	2	1
3. Short time per item	1	3	1	1	1	?
4. Little printed space per item	2	2	2	1	3	1
5. Easy to score; no partial credits	1	2	1	1	1	3
6. Free from chance success	3	1	1	1	1	3
7. Complexity of thinking	?	?	?	1	1	1
8. Question clear, not a puzzle	2	3	1	1	1	2
9. Dependence on recall, not recognition	3	2	3	2	2	1
10. Analysis of results *	3	2	1	2	2	1

* Evidence showing why the examinee failed; types of errors or omissions.

scoring are reflections of personal experience in these activities. For many test situations, however, the advantages and disadvantages of various forms are fairly clear.

Table 19 shows that the true-false test item is ranked first in three aspects: administration, ease of scoring,[1] and short time-per-item. It is given the second rank for economy of printed space, ease of composition, and clarity; and the third, for freedom from chance success, dependence on recall rather than recognition, and analysis of results. Complexity of thinking on true-false items is usually small, though it can be made very great; hence no definite rating is given. This aspect is assigned a question mark. True-false test items are probably the best to use if:

---

[1] Another aspect, accuracy of hand scoring, was investigated by Dunlap (1938). From a rescoring of 398 Terman Group Tests of Mental Ability it was found that true-false and two-choice items had nearly twice as many errors as completion or multiple-choice items. Nearly 10 per cent of items were mis-scored on the true-false tests, a very serious error.

1. One is faced with a situation in which time is very short for composing, administering, and scoring a test
2. The test will have to be scored by clerical helpers who are ignorant of the subject
3. Complexity of thinking and recall of information are not considered as important as a wide range of information
4. Occasional chance successes and failures are not too serious

In Table 19 the completion type of test item is rated high in ease of administration and freedom from chance successes. It ranks second in ease of composition, amount of printed space, ease of scoring, dependence on recall rather than on recognition, and analysis of results. It is third in the length of time needed and in the clarity of the questions, but the complexity of thinking is indeterminate for this type. Some, who prefer the completion type to the true-false type, believe that completion items are less likely to suggest wrong answers and to encourage superficial thinking.

A type of test item which combines some of the good points of both the true-false and the completion types is the multiple-choice type. It is administered and scored about as fast as the true-false type, and it is nearly as free from chance errors as the completion type. A multiple-choice item is particularly well suited for mathematics, spelling, and vocabulary tests. It is much more widely used in standardized test forms than any other type of item.

Matching-test items are particularly well suited for the classification of facts, such as in Chemistry, Table 4 B, page 26, and Mechanical Appliances, Figure 8. They allow some analysis of results and eliminate chance successes fairly well. They are given the highest rank in Table 19 in six aspects, and the second rank in the other four aspects.

A rearrangement-test item has approximately the same good points as a matching-test item, and may allow slightly more precision in scoring. Directions for matching and for rearrangement items are not as easy to understand as the directions for some of the other types, but they can be readily grasped by the average sixth-grade pupil, or an adult who is above the lowest 8 per cent of the population in academic status.

An essay item calls for an oral or a written explanation or exposition of facts. It is one of the easiest types of item to compose, yet the hardest to score, since innumerable variations appear in the answers. It usually demands more recall and organization of ideas than any of the other types. For this reason the essay-type item is widely used in appraising problem-solving abilities of advanced students. It is also the most usual way of appraising verbal compositions. It is the only type of test in which the examinee has an opportunity to give expression to his individual style and poetic fancy. In the field of literature it can never be replaced by short-answer items. Since long essays are harder to score than short ones, a common practice in educational tests is to limit the essay to 150 words, or to a definite space. This limitation has the advantage of making the examinee think out his answer clearly before he writes it.

Another approach to the relative value of various forms of items is the effect which they may have upon students' methods of study. Meyer's (1936) report is typical of several others. He told equated groups of students to prepare for one of the following types of examinations: true-false, multiple-choice, completion, and essay. At the end of regulated study periods all groups were given all four types of examinations. He found that the students who had prepared for the essay tests showed better average scores on all forms of tests than the other students. Those who had prepared for the completion-type tests came next; while those who had prepared for the true-false and multiple-choice test had nearly the same average scores.

A number of persons have tried out tests which combine two of the forms just described. A true-false form was combined with a completion form by McClusky (1934), who asked students to correct the items which they had marked false by writing additional facts on the test form. This procedure makes the test much longer in administration and scoring than a true-false test, but it may give a clearer picture of a student's knowledge of the subject. To a limited extent it is thought to avoid the retention of false information from false items.

Other tests combine the multiple-choice and completion forms. Curtis (1928) gave his students four choices for each item and also a blank in which a still more appropriate answer might be written. This combination increases the time needed for administering and scoring the tests by a considerable amount, depending upon the particular questions and students. It has the advantage of making the students think harder to give a good answer than does the multiple-choice form. It is also of use to those who wish to revise multiple-choice test items, for students often write somewhat plausible wrong answers which, if included as given choices, would make the test item more discriminating than before. This source of alternative choices is important since it is usually very difficult to find three wrong choices which are enough like the right choice to make the discrimination sufficiently difficult, and which are not so close to the right answer as to be considered synonymous by some authorities.

A following-directions test item, page 299, is a variety of multiple-choice type, in which a number of qualifying clauses are inserted in order to make a judgment more difficult.

Another very interesting combination is made by directing examinees to read a statement, and then to answer short questions about it. The questions may be true-false or multiple-choice items as:

Place a plus sign (+) in front of each item that supports the first statement, and a zero (0) in front of each item which does not.

The Helmholtz theory of color vision is supported by the fact that:

______ Some persons seem unable to see red as well as normal persons.

______ Pure red, green, and blue lights give white if mixed in the proper proportions.

______ The complement of yellow is blue.

______ The periphery of the retina is insensitive to faint colors.

______ The rods have lower thresholds than the cones.

______ Persons with normal vision can match all colors by mixing only three colors in various proportions.

This type of question is being extensively used, for it allows the examiner to appraise complex inferences while using mechanical scoring techniques.

III. RELATIVE TIME ALLOWANCES

Ratings similar to those in Table 19 on time-per-item have been determined somewhat more precisely for special cases by the study of the periods actually needed by various persons. When various types of items are constructed to have nearly the same content, as reported by Ruch and Stoddard (1927), then the relative time-per-item can be determined with considerable accuracy. Completion items and items which had seven choices were both answered at the rate of about 4 per minute; five- and three-choice items at the rate of 5 per minute, and two-choice items at 6 per minute. These figures apply to particular historical information tests given in high school.

The writer has found that, in general, easy information items show very small differences in time between completion and true-false forms. Hard information items are answered much more quickly in the true-false than in the completion form, sometimes five times as rapidly. When complexity of reasoning is an important factor in success, then the time needed per item increases greatly. When the various forms of items are not constructed so as to be nearly equivalent in content, it is not possible to predict any typical relationship.

The fact that two-choice or true-false items are answered more rapidly than the other forms of items makes it possible to use a greater number of two-choice items in a given period. Thus the true-false tests may make up for their chance errors of measurement by including a larger number of items (see Table 21, Column 8).

IV. THE CONSTRUCTION OF INFORMAL APPRAISALS

The first task for one who is to construct an informal examination is to make oneself familiar with the basic considerations in this chapter and the last. The next task is to make a decision concerning the main goal of the test. If the goal is to measure information regarding a particular course of study or occupation, then a master list of the important facts is required. If it is to measure ability to interpret data or to solve problems, then a set of typical and important tables or problems must be collected. If the goal is to appraise skill in exposition or style of expression, the examiner must find topics with which the examinee is familiar. If the goal is to appraise attitudes, initiative, co-operation, or other character traits, then the examiner should find situations for observing a person making decisions about his own behavior when he is not aware that he is being observed.

Once the goal of an examination is clearly defined, then the type of examination should be selected. For measuring information alone, short-answer tests have often proved more effective than essay tests because they cover a large number of items in a short time. For measuring interpretation of data and expression, essay tests are probably equal to or more effective than short-answer items, but many fine complex interpretation tests are now appearing in short-item form. Some of these for verbal material are shown in Table 4 and for mechanical thinking in Table 64. For appraising character traits check lists or ratings are usually made although more thorough studies actually measure such traits as honesty by test situations (see Chapter 18).

The following rules have proved helpful in the construction or revision of nearly all types of test items.

V. RULES FOR ITEM CONSTRUCTION

1. Use positive statements.

This rule may be broken occasionally, but negative statements usually lead to confusing double negatives, particularly in true-false statements. Thus the item:

T F An IQ of 70 is usually not exceeded by the lowest 1 per cent of a large random population.

is true, but it is less likely to be misread if it is revised, thus:

T F An IQ of 70 is usually exceeded by the highest 98 per cent of a large random population.

2. Avoid the unqualified use of words which have two or more meanings.

This rule is very hard to follow because one's own mental set gives only one interpretation of a statement at a time. Thus the essay item:

Describe the principal factors in mental test success.

allows at least two interpretations. Some persons interpret this to mean: Describe the main factors which allow an individual to succeed on a battery of tests. Others might indicate the factors which make a test successful on the market or in a particular survey. The item should certainly be revised to distinguish between these two interpretations.

3. Incorporate only one independent idea in a question.

This rule is particularly important in short-answer items. Thus the item:

T F A correlation coefficient may show the amount of coincidence of ranks on two tests and the reliability of the tests.

has a first clause which is true, and a second clause which is false or incomplete as it stands. It should be made into two items, one including the first clause, and the other somewhat as follows:

T F The reliability of a test may be shown by a correlation coefficient based on two trials of the test on the same group of persons given on the same day.

4. Avoid broad generalizations of time or place.

Violations of this rule are fairly common and very serious because the better informed students will usually know of more exceptions to a generalization than the other students, thus the item:

T F The best test for selecting women machine operators was found by Miss Hayes to be a simple measure of dexterity on a peg board.

was supposed to be true according to one interpretation of the article, but according to other interpretations it was false. It should be revised to be less inclusive, thus:

T F The best test in point of administrative costs for selecting women machine operators was found by Hayes to be a simple measure of dexterity on a peg board.

Another example of too broad a generalization is:

T F The usual correction for chance success or failure in true-false tests is justified when students are told not to guess.

This item is theoretically true but there are many situations in which it would not be true practically. It should be revised to cover a more specific situation, such as:

T F The usual correction for chance success or failure on a 50-item true-false test is desirable when some students have omitted a considerable proportion of items.

5. Avoid telegraphic brevity which leaves one in doubt as to the meaning.

Thus the completion item:

The meanings of words depend upon ____________

is too brief for a simple interpretation which can be answered by a single phrase. Likewise, the fourth-grade history item:

Lincoln's policy toward the slaves was one of ________

has too many possible answers.

6. Have the timing clearly explained by the directions.

It is usually a good idea to tell persons how long they will be allowed to work. If speed is important in the score but not emphasized in the directions, some very capable persons will work rather slowly because they were not requested to work fast. There is, of course, the probability that some oversensitive persons will be unable to do their best when told to work fast, but if the time limits are clearly explained they will at least try to adjust to the situation.

7. Have clear directions and be sure that they are followed.

This rule seems trite, but there are many violations. Many examinations allow one to try several samples before beginning the test proper. The examiner can then move around the room and see whether instructions are thoroughly understood. In completion or essay-type examinations the necessity for concise but complete answers should be emphasized.

8. For future use, all items should be checked statistically for difficulty, chance success, and ambiguity, and the results should be recorded on a separate card for each item. This allows one to build a file of items of known difficulty and value.

Two methods of procedure have proved effective in detecting ambiguous items: (*a*) observation of behavior and (*b*) statistical analyses. Observational methods are of great value in both performance and verbal testing evaluation. Ambiguity of scores can be detected in a performance test by watching several persons work. Thus, in assembly or manipulation tests, the observation may show that one person interpreted the task quite differently from another and solved it differently. Scores in such cases represent different skills and are therefore not comparable. The most valuable results from such situations are often not numerical scores, but descriptions of how different persons went about the work. Thus one person may assemble a lock slowly and without unnecessary movements and achieve the same time score as another person who moves fast and tries out a number of more or less random combinations before hitting the right one. The time score alone cannot be given an unambiguous interpretation. For the most adequate measurement of skills it is necessary to arrange the test situation or the directions so that both persons will use nearly the same processes (see Chapter 12).

In verbal tests, ambiguities of interpretation may often be directly observed either by taking the test and recording your own puzzling experiences, or by asking others to tell you how they interpret each item. Subjective analyses of this sort are very practical. In the evaluation of a teacher-made objective test, the teacher should, if possible, allow a period of a week or more between its construction and its evaluation. This period allows the teacher time in which partly to forget the original mental set or bias which determined the form and the selection of certain test items and to take another point of view. Sometimes, after such an evaluation, the results are startling. After a subjective analysis has been made and a test has been applied to group, a statistical analysis should be made.

VI. STATISTICAL ANALYSIS OF ITEMS

It has been found that good short-answer test items are not easy to construct. At first acquaintance, short-item types seem to be the answer to the overworked instructor's need for valid and easily constructed tests of information or skill. This belief is sadly dispelled by a few attempts to construct items. Even items which have been worked over by several persons often prove of little value. Thus, Wood (1927) found that nearly 30 per cent of the items used in an examination of law students were ineffective in discriminating the poorer from the better students. Anderson (1935) found that one-third of the items used in a test on educational psychology were more adequate for appraisal of the work of the course than was the whole test. Some of the items were merely useless for distinguishing the bright from the dull students, and some introduced errors of measurement through ambiguities.

A. Detection of Ineffective Items

After a test has been applied to a standard group, each individual item may be evaluated to see how well it discriminates among persons who are divided according to some criterion of ability. A criterion may be the total score on the test itself or some other score which is considered important. Table 20 shows an appraisal of seven true-false items by two methods which are similar in that they both

TABLE 20

Evaluation of True-False Items

Per Cents Passing, and Correlations

	ITEM						
	A	*B*	*C*	*D*	*E*	*F*	*G*
Highest third	95%	57%	96%	65%	33%	95%	85%
Middle third	65	54	94	67	54	92	54
Lowest third	50	59	97	63	68	64	56
Correlation with total score	.74	.14	.07	– .06	– .32	.30	.32

show how well a single item's results correspond to those of the criterion. One method divides the group into thirds on the basis of total test scores, and then finds the per cent of successes in each third. The other correlates success on each item with success on the total test. Item A is seen to have been passed by 95 per cent of the highest third, 65 per cent of the middle third, and 50 per cent of the lowest third, and its correlation with the total is .74. Such an item is considered to be a good item in the sense that it results in nearly the same classification of persons as the whole test. A test composed only of such items is said to have high *internal consistency.*

Item B is passed with approximately chance success by all thirds and its correlation is so low that it indicates random relationship. This item is shown by these figures to be too hard for this group of persons.

Item C is passed by nearly all of each third and its correlation is nearly zero. This item is obviously too easy to affect the relative standing of persons in that group.

Items D and E both indicate ambiguity. The figures for Item D show that all the thirds of the group had the same success, and they were all considerably above chance. There was a small negative correlation with the total. Item E was passed with less than chance success by the highest third. The middle and lower thirds succeeded much better on this item. The correlation was –.32 with the total. The most probable explanation of the results on Items D and E are that:

1. The persons in the highest and lowest thirds interpreted the items differently, or
2. The item measures some skill which is negatively related to the skills measured by the total scores.

Both Items D and E are of doubtful value in this test. Items F and G exhibit partial success in classifying the persons tested. Item F distinguishes between the lower third and the middle third, but not between the middle and upper thirds. The opposite is true for Item G. Both items have a positive but low correlation with the criterion. Although they may be used in the test, they are not so effective as Item A.

The statistical analysis of the effectiveness of items is so easily made that it should frequently be employed. It shows conclusively where chance successes and ambiguities affect the scores. Since the methods of eliminating chance successes are extremely important, they will be given special consideration.

B. Chance Success

1. Theoretical Considerations. In tests that require the examinee to choose one of two or more answers which are presented to him, there is an opportunity for success without knowledge. Thus in desperation many a student when faced by an unknown item has taken out his lucky penny and given his answer according to its performance.

If all students answer all items in true-false or multiple-choice tests, there is no simple correction for chance success which will change their relative rank order. For this reason, some examiners ask all persons to guess at the choices of which they are not sure. On the other hand, if some persons are cautious and omit the doubtful items, then corrections for chance successes usually change the rank order of the subjects. In many standardized tests, persons are urged not to guess; but, whatever directions are used, it has been found that some persons will usually guess to a considerable extent and some will not.

When there are but two choices, as in true-false items, and there are equal numbers of true and of false items, then half of the items would be guessed right on a chance basis alone. If we assume that a person's wrong items are doubtful items at which he guesses, and that he got as many doubtful ones wrong as right, his chance successes can be removed from his score by subtracting the number wrong from the number right, $(R - W)$.

When three choices are presented for each item, the mean chance success on a number of items is one in three, and the corrected score is the number right minus one-half the number wrong, $R - \frac{W}{2}$; when there are four choices, the corrected score is $R - \frac{W}{3}$, and for five choices, $R - \frac{W}{4}$, and for n choices, $R - \frac{W}{n-1}$.[2]

These corrections are probably not so effective as is often supposed, for the assumption that all wrong answers are pure guesses is doubtless unwarranted in a large proportion of cases. Moreover, in multiple-choice items the discriminations between alternatives are seldom of equal difficulty,—an assumption made in applying the formula for correction. Furthermore, the person who is lucky exceeds the mean chance success and gets credit for some of his guesses even when the correction is applied, and the person who is unlucky gets less credit than he should. This latter case is probably rare because a person's hunches and sometimes even his sheer guesses are usually a little better than chance. Often one has some information which can be used in selecting an answer, but which is not judged to be complete.

---

[2] Thurstone (1931) has given a general formula which is based on the correlation of rights and wrongs with each other and a criterion: $E = \frac{\sigma_R(r_{CR}\, r_{RW} - r_{CW})}{\sigma_W(r_{CW}\, r_{RW} - r_{CR})}$ when E is the optimum weight for wrong, R is the right answers, W is the wrong answers, and C is the criterion. This must be applied to a particular test in a particular situation, but it may give results which are widely applicable. It is undoubtedly the best way of weighting errors, but it has not been widely used.

Practically, one may decide whether or not to correct for possible chance successes and failures from empirical evidence and the intended use of the results. Let us see what happened when actual test results were corrected for chance effects.

2. THE RESULTS OF CORRECTING FOR CHANCE. A very good evaluation of the results of correcting for chance is seen in the work of Ruch and Stoddard (1925), who compared the following five types of tests in American history:

1. Recall: The American Revolution began in the year ______________.
2. Five-Choices: The American Revolution began in 1762 1775 1783 1789 1812 ______________.
3. Three-Choices: The American Revolution began in 1762 1775 1789 ______________.
4. Two-Choices: The American Revolution began in 1762 1775 ______________.
5. True-False: The American Revolution began in 1775. T F

Two forms of each type were constructed, each containing 100 items, and were then applied to large groups of high-school students. The recall type was always given first, and it was followed by one of the other types. Some of the results are shown in Table 21. The first column indicates that on the average more than twice as many items were chosen correctly as were recalled. When the theoretical correction for chance success was made, Column 2, the true-false type, showed approximately the same mean score as the recall type, but the other tests had considerably higher means. Columns 3, 4, and 5 contrast the theoretical chance success with the actual chance success. Column 5 shows that the five-choice type had the greatest excess of actual over the theoretical success. The excesses were probably due to a

TABLE 21

Effects of Correcting for Chance

Each Test Consisted of 100 Item Tests

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	MEANS		CHANCE SUCCESSES			RELIABILITY: EQUIVALENT FORMS		
	Uncorrected	*Corrected*	*Theoretical*	*Actual*	*Excess of Actual*	*Uncorrected*	*Corrected*	*18.7 Min. Each Uncorrected*[3]
Recall	23.0	23.0	00.0	00.0	00.0	.896	.896	.896
5-Choice	50.0	40.8	15.4[1]	27.0[2]	11.6	.886	.873	.901
3-Choice	57.0	40.6	25.7	34.0	8.3	.748	.806	.806
2-Choice	67.6	40.1	38.5	44.0	6.1	.849	.811	.902
True-False	57.7	24.8	38.5	34.7	− 3.8	.714	.578	.820

[1] Since 23 items were recalled, it is assumed that 77 items were "guessed" at. In a five-choice test one fifth of these, or 15.4, would theoretically be guessed correctly on a pure chance basis.

[2] Found by subtracting the recall-test mean from the means of the other tests.

[3] Theoretical self-correlations when the choice tests are lengthened to fill a period equal to that needed for the recall-type test.

(Adapted from Ruch and Stoddard, 1925, Table 65. By permission of the *Journal of Educational Psychology*.)

small but important additional context furnished by each choice that was added.

The reliabilities, as shown by correlating two equivalent forms, are given in Columns 6, 7, and 8. The scores which had been corrected for chance have nearly the same reliabilities as the uncorrected scores. Column 8 indicates that all the forms would have had nearly the same self-correlations if they were constructed to fill the same period of time. Another important finding, not displayed in the table, is that all of the choice tests showed larger standard deviations when corrected for chance. This increase is due to the fact that students with lower uncorrected scores lost more by the correction than those with higher scores.

3. THE RESULTS OF TELLING STUDENTS NOT TO GUESS. In a similar investigation Ruch and Stoddard (1927) directed half of a large group of students to guess, and others not to guess when in doubt. The results are shown in part in Table 22 for 2,453 pupils in Grades VII, VIII, XI, and XII. The raw score correlations of any two equivalent forms were usually slightly higher among students who were told not to guess than among students who were told to guess. Self-reliabilities were thus improved, compare Columns 1 and 4; and corrections for chance successes were then of little value, compare Columns 4 and 5. This result indicates a slight advantage in telling students not to guess. This table also shows that some students did not guess when asked to do so, for if all students had answered every item, the correction for guessing would not change anyone's position in the group, and hence the correlations for corrected and uncorrected scores would be the same in Columns 1 and 2.

The correlations of the scores on choice tests with the recall-test scores were nearly the same for both groups of students. The corrected scores corresponded very little better than the uncorrected to recall-test scores. If one assumes that the recall type of test is the most valid form of measurement, then he must conclude that directions to guess or not

TABLE 22

Effects of " Guess " versus " Do Not Guess " Directions

TEST	RELIABILITY: 2 EQUIVALENT FORMS						VALIDITY: CORRELATION WITH RECALL TYPE						MEANS			
	"Guess"			"Do Not Guess"			"Guess"			"Do Not Guess"			"Guess"		"Not Guess"	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
	Unc.	*Corr.*	*Diff.* (2–1)	*Unc.*	*Corr.*	*Diff.* (5–4)	*Unc.*	*Corr.*	*Diff.* (8–7)	*Unc.*	*Corr.*	(11–10)	*Unc.*	*Corr.*	*Unc.*	*Corr.*
Recall	.950												27		27	
7-Response	.80	.84	.04	.89	.91	.02	.87	.87	.00	.93	.93	.00	50	41	45	40
5-Response	.86	.90	.04	.86	.88	.02	.91	.91	.00	.89	.92	.03	54	43	48	42
3-Response	.84	.86	.02	.89	.89	.00	.84	.85	.01	.85	.92	.07	62	43	54	41
2-Response	.75	.86	.11	.86	.84	– .01	.86	.86	.00	.74	.77	.03	72	43	65	45
True-False	.64	.78	*.14*	.89	.84	– *.05*	.80	.84	.04	.75	.86	.11	66	32	51	31
Mean	.78	.85	.08	.88	.87	– .01	.85	.87	.02							

NOTE: The italicized values show all differences which are 3.0 or more times their probable errors, and hence are probably "significant differences." Nearest whole numbers are used.

(Adapted from Ruch and Stoddard, 1927, pp. 289–290. By permission of the World Book Co.)

to guess had little effect upon the validity of the choice-types of tests.

Very similar results were secured by Toops (1921) using Trade Tests on college students, and by Andrew and Bird (1938) using questions in psychology courses.

These studies lead to the conclusions that:

1. The five types of test items measure different information and skills only to a small extent;
2. All types have high and similar reliabilities when work periods are the same;
3. Theoretical corrections for chance reduced total scores, increased standard deviations, and usually did not change reliability or validity;
4. Directions not to guess reduced mean raw scores from 10 to 20 per cent, reduced the mean corrected scores between 2 and 3 per cent, and increased the reliability and validity slightly;
5. The true-false and two-choice tests showed more chance results than the other tests, and hence larger corrections for chance.

Theoretical corrections for chance are not applied widely on multiple-choice items but they are usually applied to true-false or two-choice items. The reasons are as follows:

1. Corrections for chance reduce total scores, and so may give a more accurate picture of one's actual ability.
2. Corrections usually increase individual differences in raw scores.
3. The validity of a test, it is believed, may be improved by corrections for chance. On this, however, no clear evidence is available.
4. Corrections for chance will be a protection against criticism.

4. Avoidance of Chance Results. The discussion of chance leads to the conclusion that they may be small enough

to be ignored in many forms of tests, but that there are no sure ways of entirely correcting for chance errors in individual scores. There are three ways of avoiding chance successes through methods used in construction of items and tests:

1. The test may include a large number of items. This allows the *law of averages* to reduce the chance errors in the score. If 100 or more true-false items are given, the errors in scores due to chance are usually relatively smaller than when only 10 are given. A person's luck is more likely to change during a long test than during a short one.
2. The items may have four or more choices, or completion items where the possible wrong items are numerous may be used.[3] A multiple-choice test item with four choices allows only one-third as many to be guessed right as guessed wrong on the average. In tests composed of such items the correction for chance is usually a very small proportion of a total score.
3. Persons may be asked to indicate whether they are sure or doubtful of their answers, and then more credit may be given for sure answers. The proportions of sure answers which are correct are also indices of caution in the test situation. This index of caution has not been widely studied, but it seems to offer an interesting indication of a personality trait. West (1923), Trow (1923), and Greene 1929, 1938, found tendencies for persons to vary in cautiousness according to (*a*) the type of material, (*b*) their familiarity with it, and (*c*) their temperaments and training in exactness. In general, the more abstract material showed lower per cents of sure items correct than did the concrete material, and persons with high scores had

---

[3] A completion item may, however, involve a choice between two easily recalled alternatives as:

Plato was born ________________ Aristotle.
The Amazon River is ________________ in length than the Nile.

much higher per cents of sure items correct than did the poorly informed. There was a marked tendency for the less able students to mark nearly as many items "sure" as did the better students.

VII. EFFICIENT SCORING DEVICES

The design of an item determines both speed and accuracy of scoring the answer; hence a discussion of forms of items would not be complete without a consideration of practical economies in scoring. Ordinarily, there are two processes which must be completed to obtain a test score. One is the checking of answers with a key, the other is counting correct answers to get the total. Errors and omissions must also be counted when the total score is to be corrected for chance successes.

In essay- or completion-type items, the key consists of a list of correct answers which must be compared with those the subject has written. There do not seem to be any mechanical short cuts to scoring such tests. Scoring can be speeded up a little in completion-type tests, by having all answers placed in a column where they can be more easily compared with a key.

For true-false, multiple-choice, or matching tests, there are three scoring devices: stencils, automatic scores, and machine scorers which result in a great saving of time and in increased accuracy. The saving is greatest when large numbers of tests are to be scored. No simple rule can be laid down for effecting economies in scoring, for there are at least three variables: (1) the cost of clerical labor, (2) the cost of devising and printing automatic scoring keys, and (3) the cost of operating machines.

A. Stencil Scoring

True-false and multiple-choice items are often scored by means of a printed stencil or scoring key which is placed over or beside the test answers. The stencil allows a clerk to see quickly which answers are correct, and to mark them in some fashion.

In some agencies, where a large number of tests must be scored, the examinee is required to indicate his answers on a separate answer sheet. The marked answer sheets are run through a machine which prints the correct answers with colored ink near the given answers. The subject's answers can then be rapidly checked.

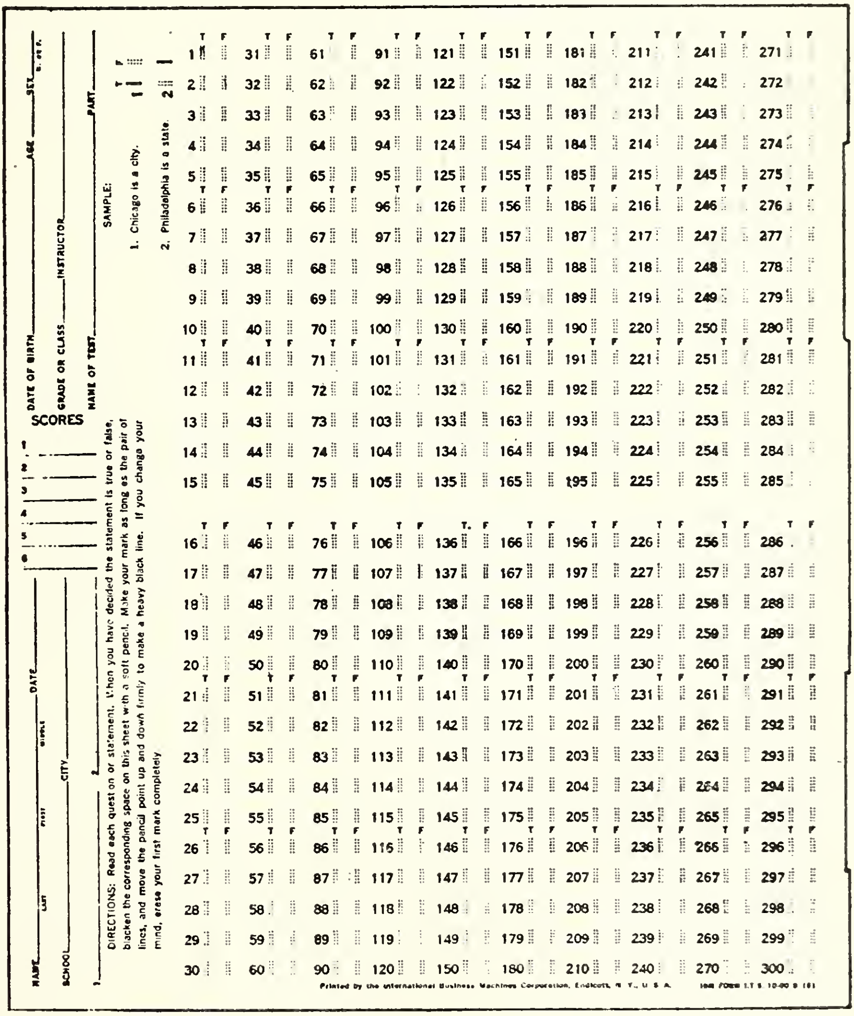
B. Automatic Scoring Devices

One of the most difficult steps of stencil scoring — checking answers against a key — is eliminated by automatic scoring. This is illustrated by a fairly large number of tests which have been printed in such a way that the position of the subject's mark on the test indicates the correctness of his answer on a key. The key is placed where the examinee cannot see it at the time he is taking the test.

Clapp and Young (1929) published a series of tests in which the key was printed on a sheet just beneath the answer sheet. The back of the answer sheet was partly covered with a thin coating of carbon. The test consists of multiple-choice items which are followed by four small numbered squares. The subject makes an X with ordinary pencil in the square which in his opinion indicates the correct choice. At the same time, his X mark is impressed on the key sheet which shows the position of the correct answers by small printed squares. To find the total correct, one simply counts the squares on the key sheet which have X's on them.

Toops (1937) used a sharp-pointed stylus for automatic scoring of the Ohio State College Entrance Examination. Test booklets and answer sheets were printed separately. The answer sheets contained rows of numbered squares. A key sheet, just beneath the answer sheet, has a square printed under each correct answer. The person being tested makes a pinhole in the square which corresponds to his choice, and simultaneously a pin hole in the key sheet. Several key sheets may be used at once when duplicate records are desired.

Greene (1938) applied the automatic scoring idea to the Michigan Non-Verbal and Vocabulary Profile Tests by having the key indicated on the back of the test answer sheet by printed circles. This method avoids a separate sheet for a key as well as special carbon surfaces and marking instru-



NAME ___ LAST ___ FIRST ___ MIDDLE ___ DATE ___ DATE OF BIRTH ___ AGE ___ SEX ___ M. or F.

SCHOOL ___ CITY ___ GRADE OR CLASS ___ INSTRUCTOR ___

1 ___ 2 ___ NAME OF TEST ___ PART ___

SCORES

1 ___
2 ___
3 ___
4 ___
5 ___
6 ___

DIRECTIONS: Read each question or statement. When you have decided the statement is true or false, blacken the corresponding space on this sheet with a soft pencil. Make your mark as long as the pair of lines, and move the pencil point up and down firmly to make a heavy black line. If you change your mind, erase your first mark completely.

SAMPLE:

1. Chicago is a city. T F
2. Philadelphia is a state.

T F	T F	T F	T F	T F	T F	T F	T F	T F	T F
1	31	61	91	121	151	181	211	241	271
2	32	62	92	122	152	182	212	242	272
3	33	63	93	123	153	183	213	243	273
4	34	64	94	124	154	184	214	244	274
5	35	65	95	125	155	185	215	245	275
6	36	66	96	126	156	186	216	246	276
7	37	67	97	127	157	187	217	247	277
8	38	68	98	128	158	188	218	248	278
9	39	69	99	129	159	189	219	249	279
10	40	70	100	130	160	190	220	250	280
11	41	71	101	131	161	191	221	251	281
12	42	72	102	132	162	192	222	252	282
13	43	73	103	133	163	193	223	253	283
14	44	74	104	134	164	194	224	254	284
15	45	75	105	135	165	195	225	255	285
16	46	76	106	136	166	196	226	256	286
17	47	77	107	137	167	197	227	257	287
18	48	78	108	138	168	198	228	258	288
19	49	79	109	139	169	199	229	259	289
20	50	80	110	140	170	200	230	260	290
21	51	81	111	141	171	201	231	261	291
22	52	82	112	142	172	202	232	262	292
23	53	83	113	143	173	203	233	263	293
24	54	84	114	144	174	204	234	264	294
25	55	85	115	145	175	205	235	265	295
26	56	86	116	146	176	206	236	266	296
27	57	87	117	147	177	207	237	267	297
28	58	88	118	148	178	208	238	268	298
29	59	89	119	149	179	209	239	269	299
30	60	90	120	150	180	210	240	270	300

Printed by the International Business Machines Corporation, Endicott, N. Y., U. S. A. IBM FORM I.T.S. 10-40 B 181

FIGURE 21. Mechanically Scored Answer Sheet

(Courtesy of the International Business Machines Corporation.)

ments. The correctness of the subject's choice can be seen when the answer sheet is held to the light.

Although these automatic scoring devices do not furnish a total score they do allow scores to be counted directly from the marks made by the person who took the test. Accuracy and speed of scoring are thus greatly increased.

C. Machine Scoring

The International Business Machine (I.B.M.) Scorer (1938) uses a carefully printed sheet, such as that shown in Figure 24, upon which the person marks all his answers with a special pencil. The sheet is printed with small parallel lines showing where the pencil marks should be placed to indicate true items, false items, or multiple choices. To score this sheet, it is inserted in the machine, a lever is moved, and the total score is read from a dial. The scoring is accomplished by electrical contacts with the pencil marks. Each sheet can be scored for ten separate divisions as well as for the total. Corrections for guessing can be obtained by setting a dial on the machine. By this method 300 true-false items can be scored simultaneously. The sheets can be run through the machine as quickly as the operator can insert them and write down the scores. The operator needs little special training beyond that for clerical work.

7

FACTORIAL ANALYSES

I. ASSUMPTIONS

The previous chapters have described the main categories of behavior and ways in which particular samples of skills and information can be given numerical representation and also checked for their accuracy of representation. Another persistent problem in the study of behavior is approached in this chapter, namely, how are abilities of persons related? Answers to this question are fairly numerous and usually hypothetical. Persons are so complex that patterns of personality have not as yet been well established.

There are, however, two methods widely used for the analysis of mental organization: one, biographical; the other, statistical. The biographical method records a sequence of events which are spatially and temporally related. It has the advantage of showing apparent causal relationships and, if carefully done, it is a valuable analysis of trends. On the other hand, it usually fails to provide a basis for the most adequate comparison between persons, and it fails to give numerical analyses of what might be considered elements or forces in the patterns described. The second, or statistical method, shows the coincidence between various measures of persons. It may be used to study either simultaneous relationships or temporal sequences. Eventually, logical syntheses of such quantitative results yield a body of natural laws.

The psychologist's search for elements of behavior is much like the chemist's search for earth elements fifty years ago. The chemists secured samples of organic and inorganic compounds and subjected them to rigid tests which showed their reactions to various forces, such as gravity, electricity, pressure, and heat. On the basis of these tests the samples were classified into groups which behaved similarly. After years of research, it was found that compounds which behave in similar fashion usually have common elements which can be isolated. An element is described as a unique substance which could not be divided into other substances. Although the isolation of an element is usually an important step in its recognition, isolation is not absolutely necessary for recognition. An element may be described and measured accurately (*a*) if it is present in different amounts in a number of the samples which are available, (*b*) if it reacts to the tests differently from other elements present, and (*c*) if its reactions to the tests are the same in combination and in isolation. This sort of analysis depends upon having a large number of samples which are qualitatively and quantitatively different.

Similarly, persons may be taken as the samples to be analyzed, subjected to various tests, and classified according to similarities of responses. For the analysis of factors in test scores, two assumptions are usually made:

1. All the scores on a given test represent the same factors. This assumption is not representative of a large number of tests where the same scores represent different ability patterns, but it is representative of carefully controlled situations. Unless one is sure that a particular

score represents the same pattern in all persons at all times, the comparisons are inconclusive. There is no absolute criterion to which one can appeal. Here, as in all other sciences, one must rely upon the agreement of observers who are considered to be competent.

2. A high correlation between two tests indicates similar processes in both tests. This assumption is also likely to be false in many test situations. A high correlation simply indicates that all persons have the same relative place in each test distribution. It does not show what factors have caused them to take these positions in the group. Only in carefully controlled situations may one make this assumption.

From these considerations it is clear that a high or a low correlation is most significant when the test conditions are most rigidly controlled. Few, if any, studies of persons have appeared that allow an uncontroversial interpretation of the factors involved in a correlation, but so much work of a preliminary nature has been accomplished that no discussion of mental measurement can ignore it. This chapter gives a brief introduction to mathematical methods that have been proposed for analyzing test relationships, a field of study in which the methods are still rapidly developing.

II. FACTORS

The last forty years have witnessed a marked development in the analysis of psychological data by means of techniques which assume that the relations found between tests are due to the existence of common elements or factors. A factor is defined as some element in the test which can be distinguished from other elements. Factors may be distinguished psychologically by observation, or mathematically by showing whether persons who are high in one test are also high in another. If scores on two tests have a zero correlation with each other, they do not exhibit any factors in common. One of the main problems is the reconciliation of mathematical and psychological factors.

III. SPEARMAN'S CONTRIBUTION

Spearman (1904) searched for the simplest explanation of the relations seen in correlation matrices. A *correlation matrix* is a table which shows the correlation of each test in the battery with each of the other tests, as in Table 23. He reported that the relation shown in Table 23 could be ascribed to one general factor which was present in all five tests in various amounts. In this case he showed that every individual score of every ability or attitude which is represented in the matrix could be divided theoretically into two independent kinds of factors: one a general factor, called *g*, and the other specific factors, which vary from test to test. A number of other explanations of the relations shown in Table 23 are of course possible, but Spearman preferred this one, since it was the simplest explanation of the intercorrelations, and it corresponded to his psychological analysis of the tests. For Spearman's arguments and mathematical proof one should consult his most recent work, *Psychology through the Ages* (1938), or Guilford (1936).

TABLE 23

Correlation Matrix for Five Mental Tests

		Opposites	*Completion*	*Memory*	*Discrimination*	*Cancellation*
Opposites	1		.80	.60	.30	.30
Completion	2	.80		.48	.24	.24
Memory	3	.60	.48		.18	.18
Discrimination	4	.30	.24	.18		.09
Cancellation	5	.30	.24	.18	.09	

(Spearman, 1927, p. 74. By permission of the Macmillan Co.)

Spearman found that a *g* factor would not account for intercorrelations in all matrices, but only in those which included tests of abstract comparison, as in Table 23. When tests were included which depended principally on motor co-ordination or sensory discrimination, one general factor could not completely account for all the correlations of the matrix. It was then necessary to include *group factors* to account for the correlations. From these studies he concluded that mental organization can be most adequately described as due to three sorts of factors: a general factor, found in nearly all mental tests; group factors, found only in groups of similar tests; and specific factors, found only in one test.

A. The General Factor (*g*)

From the analyses of many batteries of tests, Spearman and his students have tentatively named some of these factors. The general, or *g*, factor is found in large amounts in tests of abstraction, whether verbal or nonverbal, and in smaller amounts in tests of perceptual discrimination. It was not found in tests of rote memory or rote learning. Speed and accuracy are interchangeable measures of *g* when complexity is held constant. The *g* factor is therefore described by Spearman as a special sort of energy which can be applied to making comparisons or drawing inferences. Some have thought this a good definition of intelligence, but Spearman has pointed out that most intelligence tests have large rote memory factors; hence they are not good measures of *g*, although some *g* is usually present.

B. Group Factors

The group factors which have been isolated by Spearman seem to fall into two general classes — namely, those due to independent abilities, and those due to variations in *g*. The ability factors are called mechanical, arithmetical, musical, logical, and psychological. The mechanical and arithmetic factors are to be found in spatial and number tests. The logical factor is independent of the *g* factor, for from Spearman's analysis it is a technique of logical comparison which is different from a general mental energy. The psychological factors, of which there were four, are described as bases of judgments used in making decisions concerning: (1) concrete objects, (2) abstract ideas, (3) moral concepts, and (4) interest and pleasures. It was found that some persons excelled in one sort of judgment and some in another, so that correlations within these fields were higher than would be expected from the presence of a *g* factor alone. A group factor of goal consistency, also called *will,* is indicated by the fact that higher correlations than can be explained by *g* alone are found in tests which demand a persistence of motive. Spearman (1927) did not find evidence of group factors for speed, attention span, sensory discrimination, motor coordination, or language. More complete research by Holzinger (1935), however, recognizes group factors for verbal ability, motor and mental speed, attention, and imagination.

The group factors which are attributable to variations in the *g* factor are called *perseveration* and *fatigue.* These are typically character traits. Perseveration is indicated by ease or difficulty in changing quickly from one activity to another. Fatigue is indicated by the fact that a person's scores in one test may decrease after a fatiguing experience more rapidly than his scores in a different test. Groups of persons having the same fatigue patterns will show higher correlations after fatigue than other groups.

Much difficulty has been experienced in distinguishing among the general, the group factors, and the specific factors. For instance, if the variety of tests in a battery is small, a *g,* or general factor, may be found common to all of them. If these same tests were included among a larger variety of tests, this *g* factor might become a group factor. Likewise, in a small battery of tests it is quite likely that a factor will appear in one test only, and hence be called a *specific factor.* But when a large number and variety of tests are used, these specific factors will often become *group* factors. It seems, therefore, that the classification into general, group, and specific factors is more dependent upon the particular battery of tests used than upon the mental organization of persons tested. There is reason to believe that Spearman's *g* factor and many *s* factors will all become group factors when a large enough battery of tests is used.

Other factorial analyses have been described by Hotelling (1933), Kelley (1928), Burt (1938), and several others using somewhat different assumptions and securing different results. The field is still a controversial one. All of these workers have devised methods for determining (*a*) which factors would account for the relationships of a given correlation matrix, (*b*) how much of each factor is used in each test, and (*c*) how much is shown in an individual's score on a test.

IV. THURSTONE'S METHOD

The general method of Thurstone (1935) will be discussed because it has been more widely used than the others. The mathematical solution of a correlation analysis should be secured from Thurstone's work, but the main assumptions and procedures will be briefly presented here.

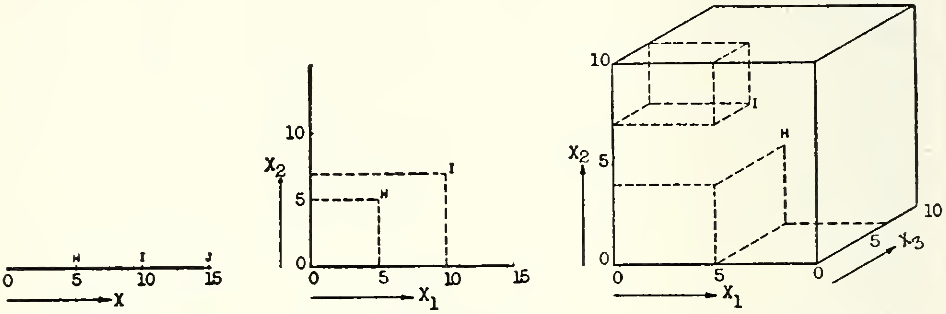
A. Analyses of a Score

Since test scores are the basic materials, some working hypothesis is needed for their analysis. Thurstone, as well as nearly all other statistical analysts, defines a test as a situation in which a person's position is determined by a number of forces.[1] The simplest possible case would be that in which only a single force is operating. This is shown in Figure 25A, where all persons in the group may be represented by points on a straight line. The direction of the force is indicated by the arrow and the amount of the force by the distance from zero to the point representing the scores of persons, H, I, J. This is the situation diagrammed in any frequency curve, ogive, or histogram. The score of H = 5x, of I = 10x, and of J = 15x. Any person's score may be represented as S = ax, where a is a certain amount of x.

---

[1] The use of the word *force* in this chapter should not be interpreted literally as a mechanical entity in all of its applications. In some instances it simply refers to a parameter in a system of measures.

A more complex score would be that which resulted from the operation of two independent forces. This is shown in Figure 25B, where the position of a person's score can be accurately described as the resultant forces x_1 and x_2. In this case H's score $= 5x_1 + 5x_2$, and I's score $= 10x_1 + 7x_2$, and any score would be $S = a_1x_1 + a_2x_2$.



A. Single Variable B. Two Variables C. Three Variables

FIGURE 25. **Graphic Representation of Independent Variables**

A still more complex test might involve three independent forces as shown in Figure 25C. Here H's score on Test A must be represented by:

$$H = 5x_1 + 4x_2 + 7x_3$$
$$I = 5x_1 + 7x_2 + 1x_3$$

and any score by:

$$S = a_1x_1 + a_2x_2 + a_3x_3$$

From these figures it is apparent that the plus signs do not indicate simple addition of similar elements, but a combination of particular amounts of forces working independently of each other. Such forces are called vectors. Their independence is shown by their different directions, since force x_1 may vary without introducing any change in x_2. When two forces operate at right angles to each other, they are called *orthogonal.* In actual space, one can distinguish simultaneously only three directions at right angles to each other. In a mental test, however, it is possible to imagine more than three factors which would be related. Hence, the score of a person on any test may be indicated by:

$$\text{Score} = S = a_1x_1 + a_2x_2 + a_3x_3 \cdots a_nx_n$$

B. Analyses of a Group of Scores on a Test

Since any score on a particular test may be considered as the resultant of the component forces, a group of scores may likewise be thought of as resulting from the same forces. A group of scores is best described by their variation from a central tendency. The standard deviation, which is commonly used as a measure of dispersion, is a convenient index for the analysis of variation of a group of scores. If only one force is effective in a test, then the standard deviation is a direct measure of the variation of the force among individuals. If, however, we assume that two, and only two, forces are operative, then the standard deviation is a resultant of both forces. If we further assume that the two forces are independent of each other, they can be diagrammed at right angles to each other as in Figure 26. There the total scores on Test A are represented by the diagonal line, and the amounts of the two component forces, x and y, by the horizontal and vertical lines. The standard deviations are shown by the points under the distribution curves. The standard deviations of the forces, x and y, bear a definite relationship to the standard deviation of Test A. According to the Pythagorean theorem, the square of the standard deviation of Test A (σ_A^2) is always the sum of the squares of the standard deviations of forces x and y:

$$\sigma_A^2 = \sigma_x^2 + \sigma_y^2$$

Because the square of the sigma is so convenient a measure in these analyses, it is called the *variance* of a group of scores. The variance of a test which has only two component factors can be completely described from the variances of those factors.

When more than two factors are present, the variance of the total score still equals the sum of the variances of all the component factors. Thurstone has shown that the total variance of any test may be analyzed into three sorts of factors, called common or group factors, specific factors, and chance factors. Common factors are defined as those which are found in at least two tests in the battery. Specific factors are those found only in one test, and chance factors are those due to random variations. To simplify calculations, the variance of total scores is made equal to one by using stand-

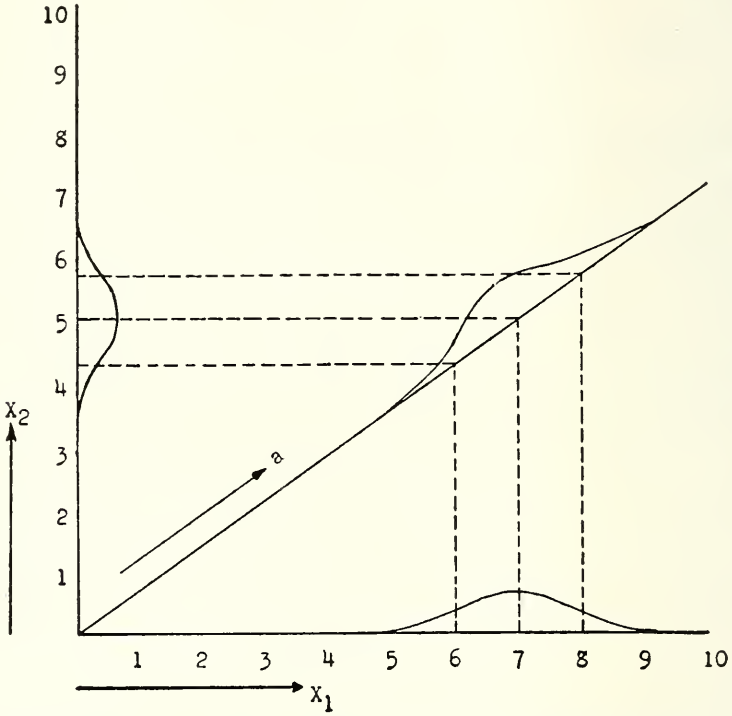


FIGURE 26. **The Dispersion of Total Scores and of Component Forces**

ard scores instead of raw scores throughout. The total variance of a test then can be indicated by:

$$\sigma_A^2 = \sigma_{A_x}{}^2 + \sigma_{A_y}{}^2 + \sigma_{A_z}{}^2 \cdot \cdot \cdot \cdot \cdot \sigma_{A_n}{}^2 + \sigma_{A_s}{}^2 + \sigma_{A_c}{}^2 = 1$$

where:

A represents total scores on Test A
A_x is the amount of variation in total scores due to factor x
x, y, z, and n are factors common to A and also at least one other test
s is a specific factor, and c is a chance factor
$\sigma_{A_x}{}^2$ is the per cent of σ_A which is due to x, since all the factors add to 1.00. It is also defined as the square of the loading of factor x in Test A

C. Common Factors

The calculation of the common factors in a test is possible from the intercorrelations between tests, because the correlation of any two tests is equal to the covariance of their common factors. Hence, a product-moment correlation can be analyzed as:

$$r_{AB} = \sigma_{A_x}\sigma_{B_x} + \sigma_{A_y}\sigma_{B_y} + \sigma_{A_z}\sigma_{B_z} \cdot \cdot \cdot \cdot \cdot \cdot \sigma_{A_n}\sigma_{B_n}$$

where:

A and B are total test scores
σ_{A_x}, σ_{A_y}, σ_{A_z}, and σ_{A_n} are the *factor loadings* of Test A in factors x, y, z, and n. The factor loading in x shows the variation in A scores which is due to variations in the x factor. It is also a correlation of Test A with factor x.
σ_{B_x}, σ_{B_y}, σ_{B_z}, and σ_{B_n} are the factor loadings of Test B in factor x, y, z, and n.

In a practical test situation, one does not usually know with any mathematical certainty how many factors are common to two or more tests. Test results, however, may be made to furnish a correlation matrix which shows the relations between tests. Thurstone (1935) described a unique solution which determines the smallest number of factors needed to account for the test correlations. His work should be consulted for a complete description. The matrix is treated as a determinant and solved by a method which assumes that all factor loadings should be positive within the tolerance of sampling errors, and that the number of zero loadings should be made as great as possible to give unambiguous results. This method gives a fairly complete solution of unique factors in a particular battery of tests, without, however, giving the factors names which represent particular forces or psychological processes. Names of processes are usually given to mathematical factors from an inspection of the tests which show the highest loadings (see Tables 65 and 106).

D. Chance and Specific Factors

The chance variations in test scores (*c*) may be measured by a correlation of two trials of the test, assuming that no systematic changes have occurred between trials. The variance due to specific factors which are not found in any other test (*s*) can be found by subtracting all the other elements in the total variance from 1.

V. LIMITATIONS AND ADVANTAGES

The procedures outlined above, and factorial procedures generally, place fairly clear limitations upon the usefulness of a factorial analysis, as follows:

1. The method should only be applied when all scores in a particular test can be shown to represent the same combinations of factors.
2. The factors which are discovered represent patterns of behavior which are independent only in the sense that they do not usually appear in similar amounts in the same persons.

3. The factors may be thought of as forces or arrangements which are found within or without the individuals. They may act against or in the same direction as other forces, not in purely additive combinations alone.
4. Group factors are, like highest common factors, the most complex factors that are common to the persons who have been tested. It is quite probable that some of these factors may consist of a number of smaller independent factors.
5. The number of common factors found in any battery of tests will vary with the items which are included and the abilities of the persons tested. Hence, even though two groups of persons are appraised by the same tests, analyses of the two groups will often yield different factor loadings.

The advantages of factorial analyses are:

1. They allow a mathematical analysis of elements or forces in tests of individuals according to specific assumptions. The assumptions can be varied to fit experimental findings.
2. The analyses show which tests require several factors and which require a few factors or only one. The selection of tests or items which are pure measures of a particular factor is thus advanced. Factorial analysis is undoubtedly one of the best methods of evaluating an item, and a great boon in the construction of analytical scales.
3. They show, with statistical precision, the types of patterns which are not found together in the same individual. From such analyses individual profiles can be constructed of factors which are distinctly independent in a particular group of persons.
4. They yield analyses of all sorts of forces, intellectual, environmental, and emotional, which are operating in the test situation. Examiners often lose sight of one variable when they are closely examining another; hence a factorial analysis is a good check on subjective analyses.

VI. TYPICAL RESULTS OF FACTOR ANALYSIS

More than two hundred reports have been published during the last few years in which batteries of tests and ratings have been analyzed by methods similar to that just described. Such reports are summarized in the later chapters which describe particular test techniques. The writer believes that a very remarkable and widespread similarity of factors has been demonstrated from the analyses of appraisals of groups which differ widely in age, education, occupation, and mental growth or deterioration.

Part II

INSTRUMENTS AND RESULTS

8

TESTS OF EARLY CHILDHOOD

I. INTRODUCTION

Interest in early stages of human development has a fairly long history which is well reviewed by Baldwin and Stecher (1924). Among the more important early contributions are *The School of Infancy* by Comenius (1628), and Rousseau's *Émile* (1762), a discussion of individual freedom. These were followed by Basedow (1770), Herbart (1898), and Froebel (1826), who described and applied liberal methods of education, and led the way for a number of treatises on infants, biographies of babies, and schools for early training. Recently four developments have given a new impetus to studies of small children in the United States:

1. The psychoanalytic movement, which has stressed the importance of early emotional maladjustments, as discussed by Freud (1925), Kanner (1935), and Thom (1922)
2. The establishment of day nurseries, partly from philanthropic motives
3. The theories of a group called behaviorists — Krasnagorski (1907), Watson (1913), and Weiss (1925) — who emphasized direct observation of behavior
4. John Dewey's emphasis since 1899 on experimental and social aspects of early education

These developments led to the gradual growth of experimental training centers, one being started at the University of Chicago in 1906, and others soon after at the University of Iowa, the Merrill-Palmer School in Detroit, Yale University, the University of Minnesota, and the University of California. Now a fairly large number of colleges and universities have well-planned centers for the study of early childhood.

The main problems for research which these centers have attacked are much the same as those encountered in studies of older persons — namely:

1. What is the mean curve of development for a particular organ or function and what deviations are found in a normal group?
2. Which sequences of development are universally found; which, occasionally?
3. To what extent are developments in various activities independent of each other? To what extent may they be explained as due to general factors? Do some inhibit others?

Such problems require accurate observations over long periods of time in order to collect significant data on growth of the same individuals. To render more accurate observations, a number of tests have been designed.

Binet included fourteen tests for children from 3 to 6 years of age in his 1905 Scale. Goddard (1910) and Terman (1916) translated these and added other tests for the same age groups. Kuhlmann (1922) developed a scale which included five items each at the 3-, 6-, 12-, 18-, and 24-month levels, and eight tests each at the 3-, 4-, and 5-year levels. During the last quarter of a century a score or more tentative scales were published, which have been reviewed by Stutsman (1931). These include observations of motor development, perceptual activities, and complex verbal adjustments. This book will not attempt to review the voluminous literature on prenatal and neonatal development. An excellent review of studies of infants' sensory and motor development is found in Brooks (1937) and in Dewey (1935). Tests for infants generally include age groups from 3 to 18 months, and tests of preschool children extend from 18 to 60 months. There is no sharp line between these two, but standardized tests for preschool children are more numerous.

II. DESCRIPTION OF SCALES

There is a marked trend in the construction of preschool tests toward analytical procedures, but as yet conclusive analyses of factors are lacking. This deficiency makes it necessary to limit the descriptions of tests to brief accounts of available testing and scoring procedures. The scales which are described are included for their wide usage, or their demonstration of interesting techniques. A number of other important scales in this rapidly developing field are reviewed in Buros (1938). The following descriptions are general in nature. More thorough analyses of the mental processes involved in successful performances will be found in some of the standard manuals.

A. Gesell and Thompson (1938) Norms of Early Growth, 4 to 56 Weeks

The work of Gesell and his aides at the Yale Institute of Human Relations, carried on since 1923, is by far the most detailed and voluminous of the studies of infant development. *Norms of Early Growth,* a practical manual by Gesell and Thompson, is a companion volume to the 1934 publication: *Infant Behavior, Its Genesis and Growth.* From 1927 to 1931, 107 infants were studied by monthly examinations and by elaborate appraisals of social and physical environments. The infants were selected from homes of middle socio-economic status, with parents of Northern European extraction. Birth history, gestation period, and physical status were held within specified limits to give the group further homogeneity, which was considered desirable in that it results in less variability of norms for the group selected.

The materials used included a standard crib with a movable platform, a cup, spoon, saucer, set of cubes, rattle, bell, ring on a string, crayon and paper, pictures, a box, and a ball. These articles have minute specifications of size and shape, color, weight, and texture to insure standard testing conditions. Similar articles have been used by many other investigators.

The reactions of the 107 infants to standard presentations of these articles were recorded on printed sheets, and also on twelve thousand feet of film, and in a large atlas. Printed norms show the per cents of each age group that responded in a particular manner to a particular situation. There are, for example, 84 items listed under behavior in response to the bell, 125 items refer to responses to one or more cubes, and 48 to the presentation with a cup. Table 24 and Figure 27 give the norms for cup behavior.

Three sorts of items appear in this table: increasing, decreasing, and focal items. The increasing items are those which show larger per cents with increasing age, as in Nos. 1, 6, 7, and 16. The decreasing items show smaller per cents with age, as in Nos. 2 and 8. The focal items at first show larger per cents with increasing age, but later show smaller per cents, as in Nos. 3, 25, and 30. From tables such as these, the "critical age" for each item was determined. For increasing items, the critical age is the first age at which an item is passed by 50 per cent of the group; for focal items, the critical age is the age of the maximum percentage.

The examination of an infant results in a record of the observed responses. For convenience, a record sheet with the items printed in order of average development is provided. A plus sign (+) is placed after items which are observed, and a minus sign (−) for items which do not appear.

TABLE 24

Cup Behavior

12 weeks to 36 weeks

Cp.	*Behavior Items*	*12*	*16*	*20*	*24*	*28*	*32*	*36*
1.	Regards immediately	81	89	97	100	100	100	100
2.	Regards momentarily	37	22	9	3	—	—	—
3.	Regards recurrently	44	68	28	21	4	18	15
4.	Regards prolongedly (n.m.p.) *	56	69					
5.	Regards prolongedly	37	57	47	3	—	—	—
6.	Regards predominantly	73	93	94	100	100	100	100
7.	Regards consistently	0	5	38	66	100	100	100
8.	Shifts regard	80	73	31	23	10	25	23
9.	Shifts regard to surroundings	40	33	6	3	—	4	15
10.	Shifts regard to hand	47	45	13	7	—	—	—
11.	Shifts regard from cup to hand	20	28	15	3			
12.	Arm increases activity (s.p. or n.m.p.) *	75	79	78	93	100	100	100
13.	Brings hand to mouth (s.p. or n.m.p.)	50	21					
14.	Hands active on table top (s.p. or n.m.p.)	36	67					
15.	Approaches (n.m.p.)	44	79					
16.	Approaches	6	25	72	91	100	100	100
17.	Approaches promptly (n.m.p.)	25	55					
18.	Approaches promptly	6	13	44	81	96	100	100
19.	Approaches after delay (n.m.p.)	25	23					
20.	Approaches with both hands	6	11	34	41	69	50	50
21.	Approaches handle first	0	5	25	38	56	64	58
22.	Contacts (n.m.p.)	44	67					
23.	Contacts	6	15	69	91	100	100	100
24.	Dislodges on contact (n.m.p.)	25	52					
25.	Dislodges on contact		9	53	50	38	29	8
26.	Grasps			13	52	85	100	100
27.	Grasps with both hands (n.m.p. or s.p.)		5	22	53	52	36	42
28.	Grasps with both hands			3	28	33	32	31
29.	Grasps with one hand			9	35	52	75	69
30.	Manipulates with hands encircling cup			6	24	56	50	42
31.	Manipulates grasping by rim			0	0	45	64	62
32.	Manipulates grasping by handle			6	31	59	92	81
33.	Pushes or hits		14	31	30	41	3	4
34.	Pushes or drags cup			19	24	33	50	31
35.	Bangs on table top		—	3	6	37	36	58
36.	Turns cup over on table top			6	26	14	17	8
37.	Lifts cup			6	45	82	100	100
38.	Lifts by handle			6	35	59	79	81
39.	Brings to mouth			3	24	63	60	66
40.	Manipulates above table top			0	21	67	86	89
41.	Manipulates initially above table top		—	3	3	26	18	35
42.	Holds with both hands			3	35	63	46	46
43.	Transfers		—	—	—	19	43	42
44.	Turns cup right side up			6	3	56	71	62
45.	Rotates		—	3	—	3	21	31
46.	Drops			6	38	63	61	42
47.	Drops and resecures		—	—	7	15	39	19
48.	Fusses	6	7	12	27	7	18	11

* n.m.p. = near median position. s.p. = standard position.

(By permission of Gesell and Thompson (1938, p. 109) and the Macmillan Co.)

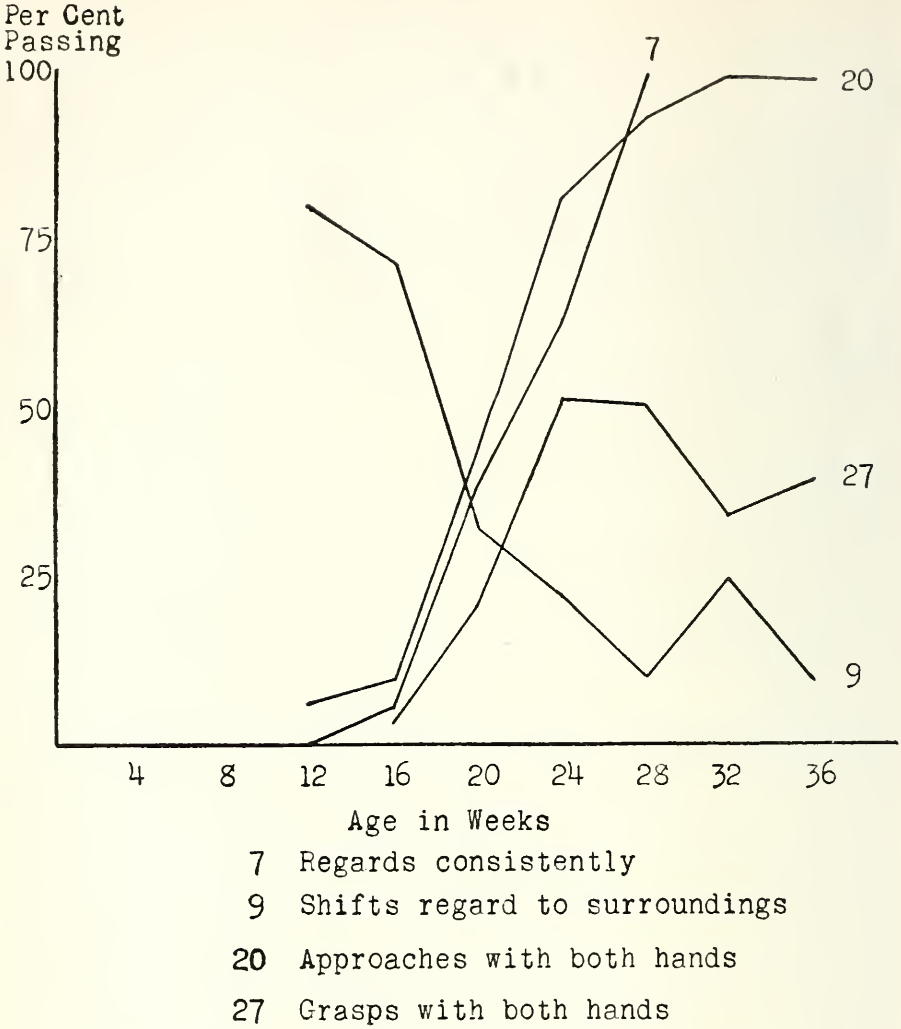


FIGURE 27. **Growth of Cup Behavior**

(Drawn from Gesell and Thompson, 1938.)

To aid in scoring, the record sheet also contains the critical age of each item, stated in weeks. A sample record sheet for postural development is shown in Table 25.

An "analysis of status" is made from this record sheet by inspecting it to find the critical age which the increasing items show by a change from plus to minus. The critical age for the child whose record is shown in Table 25 was approximately 28 weeks. A similar inspection of the de-

TABLE 25

Postural Behavior

Name: Infant G44 Date: 12-3- Age: 28 weeks

	I — Increasing Items	*Critical age*		*D — Decreasing Items*	*Critical age*		*F — Focal Items*	*Critical age*	
4				St 11 Legs flexed, not extended	4.5	−			
6	Pr 7 Lifts head to Zone 2	4.5	+	Pr 2 Head rotates (placement)	6.5	−			
	Pr 1 (Ventral suspension) Head compensates	5.5		St 1 Head sags	6.5	−			
	Pr 6 Kicks off blankets	5.5		St 4 Head sags, erect only momentarily	7	−			
				St 16 Supports no weight	7	−			
8	Pr 5 Holds head lifted sustainedly	7.5	+	Si 1 Head sags	9	−	Si 1 Head bobbingly erect	6–12	
				Su 43 Legs extend briefly	10	−			
				Su 51 Legs flexed, lifts and lowers	11	−			
				Su 22 One arm extended	11	−			
12	St 8 Head set forward or erect	9	+	Si 8 Back uniformly rounded	12	−	Pr 26 Rests only on knees, abd., chest, forearm	8–12	−
	Su 12 Arms symmetrical	11		Su 9 Arms prominently in t-n-r-position	14	−	Su 59 Progresses backward (r)	12–16	−
	Pr 8 Lifts head to Zone 3		−	Su 1 Head predominantly rotated	14	−	Pr 19 Legs flexed, outwardly rotated	8–16	−
				Pr 18 Hips raised	14	−			
16	Su 13 Arms prominently symmetrical	14		SiP 1 Head lags	19	−	Si 14 Head set forward	8–12	−
	Su 5 Head predominantly in midposition	15					Su 34 Hands in contact, arms flexed	16–20	?
	Pr 14 Lifts hand	16	+						
	Si 5 Head steadily erect	16	+						
20	SiC 2 Head erect and steady	18	+	Su 16 Arms flexed	22		Su 36 Hands engage at distance from chest	20	−
	Ra 51 Rolls to side	18		Si 21 Falls forward	23	−	Rd 48 Rolls to side	16–24	?
	Si 6 Head erect when leans forward	20	+				Pr 28 Rests only moment on abd. and chest	20–28	+
	Pr 13 Arms extended	20							
24	Pr 24 Rests on hands	23	+	SiC 4 Body slumps to side	26		Su 57 Arches back	24	−
	Si 29 Sits unsupported	23	+	Su 42 Legs flexed, outwardly rotated	27	+	Su 58 Bounces hips	24	−
	SiC 6 Body erect	25		Su 39 Legs predominantly flexed	28	+	Si 15 Sits only moment, leaning forward	24	+
							Pr 33 Rolls to side or supine	24–36	−
28	Pr 9 Lifts head to Zone 4	27	−	Pr 11 Arms flexed	30	−	Si 12 Body erect momentarily or less than 1 min.	28	−
	Si 11 Body erect	27	−				Si 28 Sits leaning forward	24–32	+
	St 20 Supports large frac. of wt. more momen.	27	+				St 35 Bounces	28	−
	SiP 13 Assists Ex. by pulling self forward	28	+						
32	St 21 Supports entire weight	30	+	Si 24 Leans forward passively	33	+	Si 23 Falls backward	32	−
	Si 17 Sits one minute or more	31	−				Su 38 Pulls foot to mouth (r)	28–32	−
	Pr 38 Pivots	31	−						
36	Si 32 Erects self from leaning forward	33	−	Si 20 Falls	37	+	Pr 30 Rests only on thigh, lower abd., hands	28–44	−
	Si 13 Body erect one minute or more	34	−				Si 27 Sits unsteadily	24–36	−
	Si 18 Sits for ten minutes or more	35	−				St 27 Stands on toes	36	−
	Su 62 Rolls prone, or attains sit'g, slt. assist.	36	+				St 15 Hips flexed, legs held extended	32–40	−
40	Si 19 Sits for indefinite period	37	−				St 37 (Hands supported) Balance inadequate	40–44	−
	Si 33 Turns to side and maintains balance	38	−				Pr 39 Regresses	40	−
	Su 63 Rolls to prone or attains sitting	38	+						
	Pr 37 Flexes leg, drawing up knee	39	−						
44	Pr 31 Assumes creeping position	41	−				Pr 32 Assumes quadrupedal position	44	
	Pr 44 Pushes upward and backward to sitting	43	−						
	Pr 40 Progresses	43	−						
	St 33 Lifts foot while supporting entire wt.	42	−						
48	St 46 Cruises or walks using support	45					CC 55 Pivots	48	
	Pr 42 Creeps	45					P-Bo 28 Turns to side, pivots, or creeps	48	
	Si 38 Pulls to standing	46					St 48 Walks only when both hands supported	48	
	St 45 Lowers self using support	46							
52	St 47 Walks using support	49					R-S 36 Turns or pivots	52	
							St 23 Feet apart 4 inches or more	52	
56	St 40 Stands independent, without support	54							
	Si 41 Attains standing independently (r)	56							
	Stc 4 Surmounts fourth tread	56							

(Gesell and Thompson, 1938, p. 211. By permission of the Macmillan Co.)

creasing items shows that the division between minus and plus signs is approximately 24 weeks. The focal items are plus at 20, 24, and 28 weeks. From these records of status the child's "maturity age" for postural development is calculated by a rough averaging process. In this example, it is calculated as 28 weeks by the authors. The more significant deviations are listed for special consideration in each case.

Similar records are made in five fields of behavior which are considered to be basic and somewhat independent. The final summary indicates status to the nearest lunar month in each field. Plus and minus signs are used to indicate small deviations. Thus, an illustrative case described by Gesell and Thompson (1938, p. 216) had the following scores (see Figure 28 A):

Postural behavior	28— weeks
Prehensory behavior	28+ weeks
Perceptual behavior	28 weeks
Adaptive behavior	28 weeks
Language behavior	32 weeks

The authors state that it is never desirable to average these scores, because this would lead to over-simplification and a disregard for careful analysis. A final summary should include (*a*) comments on the separate determinations, (*a*) their relationships, and (*c*) an appraisal of the importance of the deviations which have been noted. The authors point out that the functions are intimately related to each other, in that the youngster behaves and grows up as a total individual unit. A number of complications are listed which may change patterns of growth. Hence diagnosis and prognosis must be made from a thorough clinical study which considers all contributing factors.

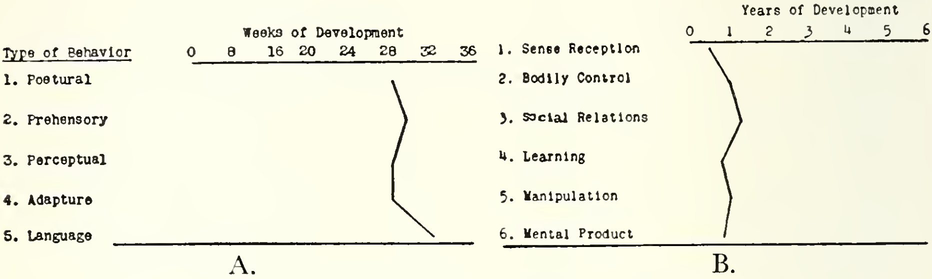


FIGURE 28. **Individual Profile**

(A. Drawn from data of Gesell and Thompson, 1938. By permission of The Macmillan Co. B. Drawn from data of Buhler and Hetzer, 1935.)

The validity of the testing schedule is thought to be high by the authors because of the way in which the schedule was developed. The manner of the selection of infants, the allocation of items into five fields of behavior, and the determination of a "critical age" for each item on the basis of percentage of infants who demonstrate the item have all been made with great care. Neither systematic errors in applying or summarizing the scale, nor random errors in scoring have as yet been reported. The training of the examiner is very important in securing comparable results.

Along with these measures of behavior, Gesell and Thompson (1938) list 14 direct measurements which are included in nearly all careful anthropometric studies:

1. Length from soles of feet to vertex (total length)
2. Length from soles of feet to suprasternal notch
3. Length from soles of feet to pubes
4. Biacromial diameter
5. Thorax diameter
6. Biscristal diameter
7. Head circumference
8. Thorax circumference
9. Weight
10. Number of erupted teeth
11. Head-neck length
12. Body length
13. Eye color
14. Hair color

From these direct measurements fifteen indices are found by dividing various measures by other measures. These specifications for measurement, described by Dawson (1936), were agreed upon at the 14th International Congress of Prehistoric Anthropology and Archaeology. Measures of infants are taken in the recumbent position, using a measuring board. Older children and adults are measured in the erect position.

B. Gesell's Development Schedules, 0 to 60 Months

A scale similar to that just described but having less detail is that of Gesell (1925). It is a schedule of observations for the neonate and for children at exactly the following ages: 4, 6, 12, 18, 24, 36, 48, and 60 months. Items were classified into Motor, Language, Adaptive, and Personal-Social Behavior. Each level has approximately 45 items, and the level was taken as an indication of his developmental status. Gesell vigorously objected to combining scores to give a general Mental Age, and emphasized the necessity for careful analysis and interrelation of observed patterns.

C. Bühler and Hetzer (1935), Birth to 6 Years

Bühler and her associates at the Psychological Institute at the University of Vienna have developed a number of interesting studies of small children over a period of years. One of the results is a standardized series of tests designed to sample six dimensions of personality:

1. Sense reception and perception (27 items)
2. Bodily movement: approaching and grasping, overcoming obstacles, and bodily control (44 items)
3. Social behavior: contact, understanding commands, and language (29 items)
4. Learning: practical memory, linguistic memory, and imitation (34 items)
5. Manipulation of material: treatment of materials and perseverance (19 items)
6. Mental productivity or goal seekings: using a tool and understanding relationships (18 items)

Each of these main divisions is appraised by means of test items. Table 26 lists the items classed under Social Behavior and Mental Productivity. The numbers in the parentheses show the age level to which that item has been assigned. Assignment of items to age levels was accomplished in such a way that the mean developmental age (DA) corresponded to the mean CA of the experimental cases. For convenience in administration, items are grouped in age levels from birth to 1 year by months, from 1 year to 2 years by three-month levels, and from 2 years to 6 years by one-year levels.

To allow a quick interpretation, a profile may be prepared, as in Figure 28 B, showing a child's scores in each of the six main divisions. Statistics are not yet available to show a factorial analysis of the items in this scale or the reliability of scores. Norms are given for 690 children, many of whom came from homes in very poor sections or

TABLE 26

Lists of Items Arranged from Buhler and Hetzer, 1935

Numbers in the parentheses after each item are the assigned mental age level: (0; 2) indicates the second month, (5) represents the fifth year, between the fourth and fifth birthdays.

III. SOCIAL BEHAVIOR

A. Contact

1. Reacting positively to the human voice (0; 2)
2. Returning the glance of an adult with smiling or cooing (0; 3)
3. Reacting negatively when experimenter stops playing (0; 4)
4. Following the movements of the adults in the room (0; 5)
5. Reflecting friendly and angry facial expressions (0; 6)
6. Actively seeking contact (0; 7)
7. Taking toy away from adult (0; 8)
8. Playing peek-a-boo (0; 8)
9. Responding specifically to gestures (0; 9–10)
10. Attracting the adult's attention (0; 9–10)
11. Turning to the adult in surprise (0; 11–12)
12. Organized ball play (1; 3)
13. Turning to the adult for explanation (1; 6)
14. Social play with a watch (2)
15. Social element in fiction play (3)
16. Moral judgment of represented actions (4)
17. Understanding the rules of a game (5)
18. Competition (5)
19. Understanding rules of a game, and competition (6)

B. Understanding Commands

1. Understanding simple commands (1; 3)
2. Understanding prohibitions (1; 6)
3. Resuming an activity on demand (3)
4. Sorting 200 slips of paper upon one command (4)
5. Performing three tasks (5)
6. Drawing a border (6)

C. Language

1. Cooing (0; 3)
2. Naming objects (1; 9)
3. Talking about absent things (3)
4. Verbal formulation of plans (4)

VI. MENTAL PRODUCTIVITY OR GOAL SEEKING

A. Using a Tool

1. Pulling an object within reach with string (0; 11–12)
2. Pulling an object within reach with stick (1; 9)
3. Using a chair as a tool (3)
4. Taking ring from hook to get object (4)
5. Using handle of a matador hammer as tool (5–6)
6. Taking off a hook to get an object (6)

B. Understanding Relationships

1. Investigating the mechanism of the bell (0; 11–12)
2. Obtaining an object from behind screen (0; 11–12)
3. Reaching for a rusk in a mirror (1; 3)
4. Preferring colored figures to plain colors (1; 6)
5. Finding a rusk under one of two cubes (1; 6)
6. Recognizing a picture (1; 9)
7. Form board (3)
8. Interpreting a picture (4)
9. Completing the puppet (5)
10. Understanding the casual relationship of three pictures (6)
11. Noticing mistakes in a picture (6)
12. Connecting merchandise and shops (6)

(By permission of Farrar and Rinehart.)

from institutions. Although the test is cumbersome in its present form and directions for administering and scoring need clarifying in places, it represents an enormous amount of careful observation and emphasizes two classifications, Social Behavior and Mental Productivity, which do not seem as well appraised in other scales.

D. The Merrill-Palmer Scale, 18 to 63 Months (Stutsman, 1931

This scale of tests was constructed to supplement the 1916 Stanford-Binet Scale, by providing more nonverbal material. Tests were chosen with the following desirable qualities in mind:

1. Attractiveness to children
2. A sampling of widely varying mental traits
3. A real difference in difficulty with increasing age, so intervals of 4 months are clearly differentiated
4. Steps equally spaced in the scale
5. Universality of experience, so that effects of environment are minimized
6. Ease in administration

In a preliminary study a large number of tests was examined. These tests were either selected from those already published, or designed from the usual activities of children. The most promising items were tried out on children at the Merrill-Palmer School. The results were analyzed to show which test had the largest number of desirable qualities. Finally, a scale containing 28 tests and totaling 93 items was devised. These tests are listed in Table 27.

No statistical analysis of the processes needed for success in these tests is offered, although brief descriptive interpretations are given for each test. Observation of the series leads one to conclude that it contains much more manipulation and nonverbal material than the Stanford-Binet, and that it evaluates time more accurately. A stop watch is used in the last eleven tests. Consequently, important factors for suc-

TABLE 27

Merrill-Palmer Scale Items

A. Language Tests:

1. Action-agent association test: Twenty words from the Woodworth-Well (1911) series
2. Ten simple questions such as:
   What does a doggie say?
   What is this? (pencil) What is it for?

B. Performance Tests:

1. "See the box. Take the box from the table and put it on the chair."
2. Throwing a ball
3. Tower of blocks, "See if you can make one!"
4. "See if you can cross your feet like that."
5. "See if you can do this." (Stand on one foot)
6. Folding a paper
7. Making a block "walk"
8. Drawing of a string
9. Identification of self in mirror
10. Cutting with scissors
11. Matching colors
12. Closing fist and moving thumb
13. Opposition of thumb and finger
14. Copying a circle with a pencil
15. Copying a cross
16. Seguin form board (See Figure 42)
17. Mare and foal test
18. Manikin Test, from Pintner-Patterson (1917)
19. Picture puzzles, 1, 2, and 3
20. Decroly matching game, No. 3
21. Wallin Peg Board (1918)
22. Fitting six cubes in a box
23. Fitting nest of cubes
24. Buttons and button holes
25. Little Pink Tower, Montessori
26. Three and six cube pyramids

(Arranged from Stutsman, 1931. By permission of the World Book Co.)

cess on this test are speed and accuracy in perception of form as well as in hand and eye co-ordination.

For standardizing the tests, 631 children were selected to secure a representative sample of the Detroit areas. Approximately 60 children were used in each age group on each test. Children on the Merrill-Palmer School waiting-list and their friends made up 33 per cent of the total, 36 per cent came from public and private school kindergartens, 19 per cent came from child-care agencies, and 12 per cent from health clinics.

The tests were administered in a quiet attractive room. A parent was often present when children below 36 months of age were tested, but not at higher ages. The children in all cases were encouraged, and no efforts were spared to keep them interested and spontaneous in their reactions to the tests. In order to maintain interest, the children were allowed to choose the next test which they were to try, by selecting the brightly colored box which they wished to have opened next.

A method of standardization, suggested by L. L. Thurstone (1925), was used. Each test was located, by interpolation if necessary, at the age when 50 per cent of the children could pass it. All tests were then ranked and numbered in order of difficulty and printed on a record sheet. This order is not rigidly followed in a standard administration of the test, but it serves as a guide as to whether or not the child has reached the probable upper limit of his performance. If the child passes more than one half of the items at one age level, he is given the next level, each level including 6 months of growth.

In summarizing this test one point of credit is given for each item passed. Omissions and refusals are credited as successes when the number ascribed to that test is below the child's total number of successes, and are classed as failures when the test number is above the child's total. Final scores are turned into mental age levels by use of a table which contains a smoothed curve for mean scores of chronological age groups. Percentile ranks and standard scores are also given for each month group from 18 to 63 months, to show a child's position in his own age group.

A correlation was found between scores on the Stanford-Binet and the Merrill-Palmer scale of .793 ± .019 for 159 children whose chronological ages ranged from 3 through 5 years. The r between CA and Merrill-Palmer test scores for 631 cases was .921 ± .004, when those who were thought to be markedly superior or feeble-minded were excluded. The overlapping of adjacent 6-month age-groups was measured by the percentage of children in the higher of two groups who fell below the median of the lower group. The median of these percentages was 21 per cent. The younger groups showed less overlapping than the older.

There are excellent detailed directions for giving and scoring the test with brief examples of scored answers. The use of the test as a means of studying personality types is explained, and a theoretical analysis of results is discussed.

E. Minnesota Preschool Tests, 1½ to 6 Years (Goodenough, 1932)

This test is divided into verbal and nonverbal parts and these may be given separately or together. To make it possible to retest the child without having the results of the second test affected by specific memories of tasks of the first test, two forms have been provided.

Both the verbal and the nonverbal series are given to the subject with verbal instructions, but the comprehension of instructions is very easy in those tests that have been classed as nonverbal. Since the author used few nonverbal tests for children under 3 years, the distinction between verbal and nonverbal tests is not considered to be useful at ages below 3 years.

A preliminary selection of tests to be used was made from a survey of previous work and from original materials. This preliminary series was applied to 100 children in each of nine half-year chronological periods. Their parents were selected in the proportion present in Minneapolis adults in order to give a cross section of occupational levels. An analysis of the children's responses was the basis for the final selection of items. Items without intrinsic interest for the children, or causing a pronounced emotional reaction, or otherwise inappropriate were eliminated. The 26 subtests listed in Table 28 were retained and arranged in a

TABLE 28

Minnesota Preschool Test

Form A

		Maximum Scores	
		Verbal	*Nonverbal*
1.	Part of body: ears ___ chin ___	2	
2.	Objects in pictures: chair ___ apple ___ house ___ flower ___	4	
3.	Naming objects: ball ___ watch ___ pencil ___ scissors ___	4	
4.	Copying drawings: circle ___ triangle ___ diamond ___		3
5.	Imitative drawing: Horiz. stroke ___ vertical cross ___		2
6.	Block building: three cube pyramid ___ six cube pyramid ___		2
7.	Response to pictures: a. nouns ___ prep or verbs ___ b. nouns ___ prep or verbs	18	
8.	Cube imitation: 1234 ___ 12342 ___ 1324 ___ 1423 ___ 14324 ___		5
9.	Command: drink ___	1	
10.	Comprehension: hungry ___ sleepy ___ house on fire ___	3	
11.	Discrimination of forms: number correct ___ number wrong ___		10
12.	Naming from memory: doll ___ pencil ___ penny ___ horse ___ shoe ___ fork ___	6	
13.	Recognition of forms: a. ___ b. ___ c. ___		3
14.	Colors: red ___ blue ___ pink ___ white ___ brown ___	5	
15.	Tracing forms: circle ___ square ___ irregular forms ___		6
16.	Picture puzzles: horse ___ goat ___ apple ___ camel ___		9
17.	Incomplete pictures: bird ___ girl ___ watch ___	14	
18.	Digit span: 2 digits ___ 3 digits ___ 4 digits ___	3	
19.	Picture puzzles: bird ___ flower ___ giraffe ___		8
20.	Paper folding		1
21.	Absurdities: a. ___ b. ___ c. ___ d. ___ e. ___	5	
22.	Mutilated pictures foot ___ finger ___		2
23.	Vocabulary: a. ___ b. ___ c. ___ d. ___ e. ___ f. ___ g. ___	7	
24.	Opposites: a. ___ b. ___ c. ___ d. ___ e. ___ f. ___	6	
25.	Clock: 8 : 10 ___ 1 : 50 ___ 12 : 00 ___ 1 : 10 ___		4
26.	Speech during examination	2	
	Scores added by ________________ Total verbal	80	
	Checked by ________________ Total Nonverbal		55

(Arranged from Goodenough, 1932. By permission of the Educational Test Bureau, Minneapolis, Minn.)

somewhat random order for a standard presentation. It was thought that this order sustained interest by variety of task and by having easy items mixed with harder items. The items were assigned various points of credit according to their difficulty. In scoring a test, the points of credit are totaled, and then changed to MA's or standard scores. The test items are arranged for convenience in administration in a large book of envelopes with instructions and materials for each test placed together.

Correlations are high between scores achieved on an equivalent form of the test given within a period ranging from one to seven days when computations are made for various chronological age groups of 6-month intervals. The correlations range from .68 to .94 for the verbal series, with a mean of .86. For the nonverbal series, the correlations range from .67 to .92, with a mean of .89. This mean indicates that prediction from scores on one test for scores on a second test within a short interval will be fairly accurate. Correction for practice effect is made by subtracting 2 C-score points from the second test, if the second test is given within the week. Goodenough (1932) states that the tests are about equally reliable at all ages within the scope of the test, from 18 months up to 6 years.

F. Goodenough Measurement of Intelligence by Drawing, 3½ to 13½ Years

Goodenough (1926) concluded, from her own observations and from a thorough study of other's research, that children's drawings could be used as an indication of intellectual development. She devised a test in which children draw a man from memory. The examiner gives the following directions to a group of children: "On these papers I want you to make a picture of a man. Make the very best picture that you can. Take your time and work very carefully. I want to see whether the boys and girls in —— school can do as well as those in other schools. Try very hard and see what good pictures you can make." Judicious praise in general terms is advised as, for example, "These drawings are fine; you boys and girls are doing very well." Suggestions from one child to others are not allowed. The test usually takes no more than ten minutes.

A man was chosen by Goodenough as the standard subject for the drawings because it is one with which all children are familiar, it has universal interest and appeal, and the clothing is more uniform than that of a woman or a child.

The children on whom the test was standardized were selected at random except for age-grade classification. They were within the normal range for *in-grade-at-age.*

The points used for scoring were chosen because they showed:

1. A regular increase in the percentage of children succeeding at successive ages
2. A rapid increase in this percentage
3. A clear differentiation between the performances of children who were of the same age but different school grades

Point credits were given for the presence of a line representing part of a man and also for correct proportion and perspective. The points are all described in detail and illustrated by specimen drawings with scores attached, as in Figure 29. It is possible to secure scores up to fifty-one points, which can be turned into mental age scores by the use of a smoothed curve. Two points is the score for the average child of 3½ years; six points is the score for a child 4½ years; and ten points is the score for a child of 5½ years, on up to a score of forty-two for a child of 13½

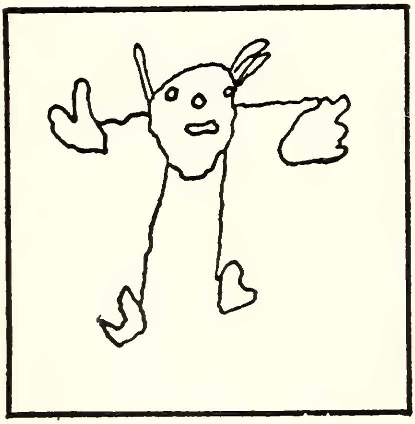


FIGURE 29

Drawing of a Man

(Goodenough, 1928. By permission of the World Book Co.)

years. Material for more complex psychological analysis of the products is given in the book, but the main use of the test is as a tentative classification of intelligence.

The use of a single sample of drawing is defended by Goodenough. She found a correlation of .937 ± .006 between two tests on successive days for 194 first grade children. In another instance, by computing the score by the split-half method, and by using the Spearman-Brown formula, a mean reliability of .77 was found for ages 5 to 10 taken separately. "The probable error of estimate of a true IQ earned on the drawing test is approximately 5.4 points at all ages from 5 to 10 years."

G. The Stanford-Binet and the Kuhlmann-Binet Tests

These two tests have well-standardized preschool sections which give the age level location of test items. These are described in Chapter 10.

H. The California Preschool Schedule

A scale known as the *California Preschool Schedule,* Macfarlane (1938), has been compiled of tests from Stutsman, Kuhlmann, Binet, Terman, and local authors. The scale is important because it is probably the most analytical of the preschool scales and because it has been used in a very thorough growth study. It extends from 15 to 84 months and has two equivalent forms. The record sheets classify the items into 10 groups, as follows:

1. Motor skills: pegs, buttons, pins, bow knot
2. Block building: tower, door, design, stair
3. Drawings: scribbling, vertical and horizontal lines, circle, triangle, diamond
4. Discrimination of form: cards, form boards
5. Discrimination of spatial relation: cover on box, in, or under
6. Discrimination of number and size: large and small blocks, count objects to 10, ½ of 6, and 8
7. Language comprehension: points to picture, action agent, opposites

8. Language facility: expressiveness, use of language, preposition, pronouns, past, plurals
9. Memory span: finds object, memory for 9 objects
10. Completion, nonverbal: mends doll, watch, pictures

This classification is descriptive of the materials and symbols rather than of the mental processes used. The scale probably does not have enough items in each class to make a reliable profile, but it does allow one to compare three types of motor skills, three types of discrimination, two of language use, and two of memory functions. Such comparisons will lead to a knowledge of the interrelation of these skills and eventually to more analytical scales.

I. Measures of Reading Readiness

A number of tests designed to indicate development which is needed by those who are starting to learn to read have recently appeared. Such tests are usually called reading readiness tests. Observations have shown that such indicators may be found:

1. By measuring the child's ability in tasks which normally are performed progressively better during the first year of school
2. By securing his intelligence test score
3. By securing teacher ratings of interest in reading
4. By measuring language ability, motor co-ordination. hearing, and visual discrimination
5. By finding how well informed the child is about his own environment
6. By learning the type of emotional adjustment or interest which he shows toward other children, adults, materials in school situation, and toward himself

The greater number of tests for reading readiness are not comprehensive enough to take in all these factors. Usually the limitations of space, time, and money reduce them to paper and pencil tests sampling a few indicators. The

Metropolitan Readiness Tests, the Monroe Reading Aptitude Tests, and the Betts Ready-to-Read Tests are typical.

1. Metropolitan Readiness Test. Although the Metropolitan Readiness Test has been designed as a group test, it is suggested that the group be small, preferably with fewer than ten children. It may be given individually if necessary, though of course the same standardized directions are to be followed. Hildreth and Griffiths and Orleans (1933) described their test as follows:

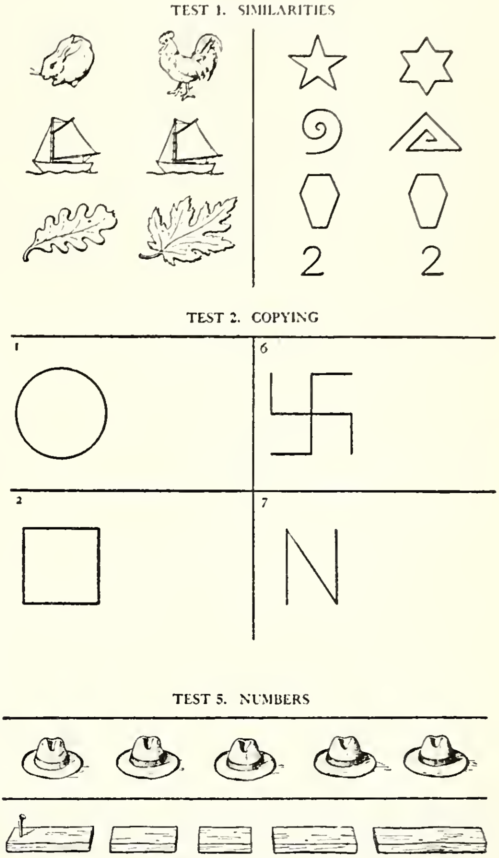


FIGURE 30

Samples from the Metropolitan Readiness Tests

(Hildreth, Griffiths, and Orleans, 1933. By permission of the World Book Co.)

> The test consists of six parts. The first is a test of perception, involving recognition of similarities. It consists of 23 items ranged in order of difficulty and including both pictorial material and symbols (see Figure 30).
>
> Test 2 is a second perception test, involving copying of 11 figures. This type of test has proved to be highly diagnostic of mental maturity in young children, and several of the items are comparable with those contained in the Binet series. The factor of reversals, which enters into a number of the items, has been found to be correlated with lack of experience and with immaturity of perceptual abilities in young children.
>
> Tests 3 and 4 are designed as measures of vocabulary. They consist respectively of 19 and 15 sets of rows of four pictures

each. In Test 3 the child is to select the picture that illustrates the word the examiner names. This is a test of understanding or comprehension of language – not a test of the child's language usage. Test 4 is similar in organization, but requires the child to comprehend phrases and sentences instead of individual words. The extra conversation, which is really not absolutely necessary for the location of the right picture by the child, is added to make the test one of more sustained attention comparable to the attention span required to listen to stories and the like in the beginning work in reading.

Test 5 measures number knowledge. By means of 40 items it measures achievement in number vocabulary, counting ordinal numbers, recognition of written numbers, writing numbers, interpreting number symbols, the meaning of number terms, the meaning of fractional parts, recognition of forms, telling time, the use of numbers in simple problems. By using the picture form of material and certain checking devices, the varied aspects of number knowledge can be very satisfactorily explored.

Test 6 appraises common knowledge by means of a multiple-choice picture test of 16 items. The child is required to select from a row of four pictures the one that satisfies the examiner's description. The test is short, but gives satisfactory evidence of variability within a group at first-grade entrance.

Centile norms for each half year from 5½ to 8 years are furnished for total scores. These show slightly larger improvements by the advanced than the retarded children during this period. The average child improves from a total sixty-four at 5 years and 9 months, to seventy-eight at 7 years and 9 months. This increase of fourteen points, or approximately 23 per cent, is equal to .67 of the sigma of the 5-years and 9-month group.

2. MONROE (1935) READING APTITUDE TESTS. In the Monroe Reading Aptitude Tests there are included items similar to those in the Metropolitan Readiness Tests and also measures of immediate memory spans, articulation, auditory discrimination, speed and accuracy of hand and eye co-ordination, and lateral dominance. Seventeen tests are grouped into six sections each of which has a separate score. A profile of six ability patterns is therefore available:

A. *Visual Tests*

1. Memory of Orientation of Forms: The child must observe a larger diagram which the teacher exposes and then draw a line around the one, of two small pictures in his booklet, which is like it. The two small pictures are lateral or vertical opposites of each other. The test was designed to detect perceptual reversals. Twelve pairs of pictures are used.
2. Ocular-Motor Control and Attention: This test, which is very similar to the McQuarrie Mechanical Aptitude line tracing test, requires one to indicate the path from a picture of a boy to one of three small pictures of houses. Nine diagrams of increasing complexity are used.
3. Memory of Forms: This test asks a child to draw a set of four small pictures immediately after they have been exposed for ten seconds. Four sets of pictures are used.

B. *Motor Tests*

1. Speed: The child is asked to place dots in ¼-inch printed circles for sixty seconds.
2. Steadiness: The task is to draw a pencil line on a row of dots and dashes, without time limits.
3. Writing Name: The child is asked to write his name with each hand.

C. *Auditory Tests*

1. Word-discrimination: A small picture is presented and three words are pronounced. Thus, for the picture of a small sailboat, the words *beet, boat,* and *boot* are spoken. The task is to indicate which of the three words is the name of the picture. In all, nine pictures are presented.
2. Sound-blending: Three small pictures are presented at a time, and the examiner pronounces the name of one of them, separating vowel and consonant sounds. The task is to blend the sounds into a word and to circle the picture which was named.
3. Auditory Memory: A short story is read aloud by the examiner, and the child is asked to "tell what the story was about." Twenty-two ideas may be given separate credits.

D. *Articulation Tests*

1. Reproduction: The child is asked to pronounce a series of 24 words or phrases, ranging in difficulty from *baby* to *transcontinental.*

2. Speed: Here the directions are, "I want to see how quickly you can talk. When I say go, say *banana, banana, banana, banana,* as quickly as you can. Keep on saying it till I say stop. Ready, Go!" Fifteen seconds are allowed. Similar trials are given, using *long ago* and *take a bite.*

E. *Language Tests*

1. Vocabulary: Twelve rows of three small pictures each are presented. The task is to draw a circle around the object which is named by the examiner.
2. Classification: "Name all the animals you can think of as quickly as you can." Thirty seconds are allowed. The same procedure is used for things to eat and for toys.
3. Sentence Length: A picture of a farmer and two boys is shown while the examiner says, "Here is a pretty picture; what is the picture about?" The score is the number of words in the longest sentence or phrase used.

F. *Laterality Tests*

Score equal number of right side preferences.

1. Hand Preference: Writing, throwing a ball, pretending to comb hair, holding a bat (shoulder), threading a needle, winding thread on a spool, folding hands (uppermost thumb), and folding arms (outermost arm).
2. Eye Preference: The child holds the large end of the paper cone to his face and is asked to look at three objects about the room. The eye used for sighting is noted. The eye used to peep through a small hole in a piece of cardboard which the child holds before his face is also noted.
3. Foot Preference: Hopping, kicking an imaginary football, and climbing upon a low chair.

Ten of the items may be administered to small groups, and seven must be given to individuals. Norms for 434 children from 5½ to 8½ years are given in smoothed percentile curves. The odd-even reliability of the entire test was found to be .87, but that of the separate sections was not given. The correlation between total Monroe scores and reading achievement one year later, as shown by a combined score on Gray's Oral Paragraphs and the Iota Word Test, was .75 among 85 Grade I children. The separate sections of the Monroe test all correlated between .50 and .60 with the combined score, with the exception of .18 for right-side dominance. Monroe suggested remedial procedures for aiding children with special difficulties.

3. BETTS (1936) READY-TO-READ TESTS. All of the Betts Ready-to-Read Tests are given individually and require more complicated apparatus than the others mentioned above. They do not include special measures of comprehension but do include four types of measures, as follows:

A. *Visual Perception*
  1. Discrimination of letter forms
  2. Discrimination of word forms and vocabulary
B. *Auditory and Articulation Tests*
  1. Phonetic perception of letter groups
  2. Auditory span for sentences
  3. Auditory fusion of vowel and consonant sounds into words
  4. Repetition of sounds
  5. Acuity of hearing: Number read in a low voice at 20 feet
C. *Visual Mechanism: Tested by a stereoscope*
  1. Acuity of each eye and of both
  2. Binocular fusion and balance
  3. Depth perception
  4. Near-sightedness
D. *Lateral Dominance: Hand, foot, and eye*

Measures of these sorts have generally been found to show low correlations with Grade I reading success because motivation, information, and reasoning ability may overcome mechanical and perceptual handicaps. On the other hand, auditory and visual handicaps should be detected at an early age to avoid unnecessary strain.

J. Summary of Description of Scales

These brief reviews are only enough to show that the items included in the various scales employ a wide variety of test situations. A clear understanding of test results comes only from a large number of direct observations of actual tests in progress and a thorough study of child development.

Two methods of selecting items have been illustrated here. In one method, stimulus situations were secured which were thought to indicate a factor of intelligence or to predict reading ability. In the other method, shown in Gesell's and in Buhler's work, there was at first a wide random selection of stimulus situations intended to give a fairly complete picture of behavior. Later, items were grouped into classes which seemed to show related forms of behavior. Selection by means of careful statistical analyses of relationship has not yet been accomplished in any of these scales, but great efforts are now being made in this direction. The scaling of items in all of these tests is achieved by converting raw scores into mental ages.

III. TIME SAMPLING

A time-sampling procedure secures indications of abilities or interests by recording the proportion of time which is spent in various activities. Thus, Arrington (1932) reported the activities of 41 children, 18 to 48 months of age, during 24 five-minute periods. She found that the average 2-year-old spent about 75 per cent of his time manipulating materials, 12 per cent in walking, running, and jumping, and 12 per cent in non-overt activities such as inactive watching. Van Alstyne (1932) studied 112 children, 2 to 5 years old, for 15 periods of forty-five minutes each. Her children spent nearly 98 per cent of the time in play with materials. Interest in other children increased with age, from 10 per cent of active co-operation shown by the 2-year-olds, to 30 per cent shown by the 5-year-olds. Constructive activity gradually replaced mere manipulation, for only 6 per cent of the 2-year-olds showed dramatization of construction, while 75 per cent of the 4-year-olds and 80 per cent of the 5-year-olds exhibited these patterns.

The time-sampling methods have usually yielded very high agreements between trained observers who were watching the same persons. Arrington's (1932) observers regularly reached correlations of .96 or more between amounts of time spent in various activities. On similar data, Van Alstyne (1932) reports correlations which ranged from .89 to .99.

IV. SOME RESULTS OF MEASUREMENT

Typical results of measurements of preschool children will be discussed under three headings: usual order of development, individual profiles, and correlational studies of growth. A discussion of the measurement of growth is found in Chapter 21.

A. The Usual Order of Development

A number of observations have been made to show the order of development of behavior patterns. In the sphere of motor co-ordination the results are fairly uniform from person to person on a number of patterns. Thus Shirley (1933) reported a definite order of emergence of motor patterns in locomotion, and Halverson (1931) reported approximate ages at which a definite series of prehension patterns developed. He described the stages as follows (p. 218):

1. *No contact* includes all instances wherein infants for some reason fail even to touch the cube.
2. *Contact only* includes instances in which infants succeed in touching but not securing the cube.
3. *Primitive squeeze* — the infant thrusts the hand beyond the cube and corrals it by pulling it toward him on the table with thumb, wrist, or hand until he succeeds in squeezing it against the other hand or the body. The primitive squeeze is not a true grasp for the hand does not actually grip the cube.
4. The first form of actual grasp is the *squeeze grasp*. The hand, palm-in, approaches the cube laterally on the table to envelop it. At the moment of contact the fingers close on the cube so as to press it strongly against the heel of the palm with the thumb extended on the upper face of the cube. This grasp is clumsy and usually results in failure to hold the cube, at least

no infant has succeeded in raising the cube from the table with this grasp.

5. In the *hand grasp* the infant brings the pronated hand down paw-like fully upon the cube, curls the fingers down on the far face of the cube with the thumb paralleling the digits on the adjacent surface, and presses the cube firmly against the heel of the palm. The fingers appear to be of equal importance in grasping and the thumb seems to lack tonicity.

6. The *palm grasp* is accomplished by setting the pronated hand down on the cube so that the fingers curl over the top and far down on the further face with the thumb pointing down against the near face to oppose the fingers in forcing the cube against the palm. We have now for the first time active thumb-opposition. This new feature in the thumb repertoire of functions and the simultaneous budding into prominence of the forefinger are mainly responsible for the higher types of grasp which follow. Up to this point all digits function in holding the cube in the middle of the palm. From now on only the first three digits function prominently in grasping so that we find the cube no longer in the middle of the palm but shifted to the radial edge of the hand. A faulty palm grasp becomes a hand grasp. As a matter of fact, the hand grasp is often due to the failure of the thumb to orient itself properly for opposition to the other digits.

7. In the *superior-palm grasp* the infant sets only the radial side of the palm down on the cube with the thumb against the near side opposing the first two fingers, which are curled down on the far side. In closure, the digits press the cube against the thumb and palm.

8. The *inferior-forefinger* grasp greatly resembles the superior-palm grasp. There is the same thumb-forefinger opposition, but the digits at the end of the approach point more medianward than downward (tendencies of this change of pointing appear in the superior-palm grasp), an angle of approach which makes thumb-opposition simple, and the cube is no longer pressed against the palm. This type of grasp represents an achievement of no small degree for the infant, for here is a clear demonstration of the fact that the digits are beginning to act independently of the palm in grasping and holding. Heretofore, the infant uses the palm in grasping to make up for the shortcomings in gripping by the digits. Now he finds that he can make the necessary fine adjustments of the proper digits to secure his grasp and can maintain the delicately balanced pressure of the digits against the near and far faces of the cube to insure its being held.

We may say that the advance in grasping from the palm type to the forefinger type marks the change from a three-surface grasp to a two-surface grasp — a tri-directional pressure gives way to a di-directional pressure. The inferior-forefinger grasp is not a true finger-tip grasp, however, for the cube is still well up toward the palm so that a considerable portion of the palmar surfaces of the digits contacts with the cube.

9. The *forefinger grasp* is essentially a finger-tip grasp. The cube is well out on the tips of the first three digits (sometimes four digits) with the thumb opposed to the fingers. Up to this point, all types of grasp require that the hand come to rest on the table before the cube can be raised. Thus the table serves as a base or leverage point for lifting the hand after it grasps the cube. In the forefinger type of grasp the digits are pretty well extended, distinct flexion appearing only at the metacarpophalangeal joints. In earlier forms of grasp the digits curl about the cube.

10. The *superior-finger grasp* is exactly similar to the forefinger grasp, with the exception that the infant in grasping does not have to place any portion of his hand on the table top to aid the placement of the digits against the cube, nor does he require the table for leverage in raising the cube. The hand alights on the cube, attains its grip, and raises the cube deftly and neatly. True, the hand may touch or brush the table as the hand settles on the cube, but the presence of the table about the cube is not essential to grasping or raising it.

Similar studies also show order of growth in language patterns. Within the first few hours after birth there is, according to Lewis (1936), a period during which the cries accompanying discomfort and the sounds accompanying comfortable situations can be distinguished by careful observers. This is followed in the second or third months by *babblings*. Babblings are repetitions of comfort sounds which seem to be produced because they give pleasure to the one making them. Lewis believes that this is a rudimentary form of aesthetic experience which may profoundly affect the adults' appreciations of music and poetry. During the first three or four months there also occur imitations of adult speech in which there is only a rough similarity between what the child hears and what he produces. According to Lewis, imitation of intonation is rare unless the child is experiencing the same emotional pattern as the adult.

At about the fourth month there follows a stage which lasts from four to six months, during which very little imitation is reported by many careful observers. This marked decrease in imitation is thought to be due to the fact that during this period the child develops responses to the meaning of what he hears and that these new associations prevent responses to mere sound. There is also a development of variety in vocalization. This period is followed by one in which a large amount of definite imitation occurs of both form and intonation. For the first time *echolalia* and imitation are noted. Echolalia is a marked tendency for a child to imitate immediately what he hears in an apparently meaningless way. This is the kind of imitation that is characteristic of extremely feeble-minded children at later ages. It also gives a child important practice in speaking adult phonetic forms. Persistent echolalia may prove a hindrance to the acquisition of language as an instrument, but it is a necessary stage of development during which one gives close attention to the perception and vocal reproduction of words.

The growth of language from this point on is a matter of very rapid accumulation of phonetic forms and new concepts. A child comprehends more spoken words than he uses in his own speaking, and this doubtless lasts throughout life. Also, a 5- or 6-year-old understands written words before writing them. These processes, however, come in fairly close succession, and they are all dependent upon the concepts which the individual develops. Precise concepts lead to rapid and accurate use of words, and this precise use of words leads in turn to more complex concepts. The order of development of concepts and modes of thought is not easy to appraise, since observations and test results usually record end-results and not the mental processes which produced them.

Challenging and much challenged discussions of concept development are found in Piaget's three volumes (1926;

1929; 1930), in Stern (1914), in Koffka (1928), and in the reports of psychoanalytical investigators, as Anna Freud (1925) and Kanner (1935). These investigators have not reported standardized testing methods, but they have made detailed observations which are of great value.

The results of the most careful testing of both language and nonlanguage patterns are still far from complete, but the order of difficulty of items on standard tests is an indication of their most probable order of development in the individuals who were used to standardize the test. From Table 24, Cup Behavior, one could calculate the probability that any item in the list would precede any other item. The items which reach a maximum earlier will probably precede the items which reach a maximum later. The order of development will almost certainly be: regarding, approaching, contacting, grasping, and manipulating.

The development in social behavior, Table 26, also shows a fairly definite order in which simpler types of perception and memory precede the more complex co-ordinations and judgments.

B. Individual Profiles

Two of the scales previously described yield data for individual profiles, Figure 28 A and B. These profiles have some similarities. Gesell's Postural and Prehensory Behavior are similar to Bühler's Bodily Control and Manipulation. Gesell's Perceptual class resembles Bühler's Reception. Gesell's Adaptive class is divided between two of Bühler's classes — Learning and Mental Productivity. Gesell's Language Behavior is distributed in several of Bühler's classes, while Bühler's Social Behavior is not well represented in any division of Gesell's scale at this age level.

These comparisons illustrate the fair degree of agreement which is commonly found in subjective analyses of this sort. They also emphasize the need for research to show what patterns of behavior are most independent of each other and most significant for predicting development. One approach to these problems is through the use of correlation analyses.

A developmental age unit was used by both Gesell and Bühler for scaled comparisons. An absolute unit would probably give other valuable information. Gesell and Thompson (1938) used weeks of mean growth as the unit, while Bühler and Hetzer (1935) used months and years of mean growth.

From graphs of this sort the variation within an individual can be studied. The infant of Figure 28A shows a range of scores of only 4 weeks, with Language Development at 32 weeks and other types of development at or about 28 weeks. The infant of Figure 28B shows a range in developmental age from 6 months in Sense Reception to 15 months in Social Relations. The other four types are near his median of 10 months.

C. Correlational Studies of Growth

Correlation analyses of measures of small children at various ages have not yet given very clear results because the samples were small and the tests were usually subject to large accidental variations. A number of correlations have already been quoted along with descriptions of the scales. The following reports are all fragmentary but they do indicate some typical relationships, and they point the way for future research.

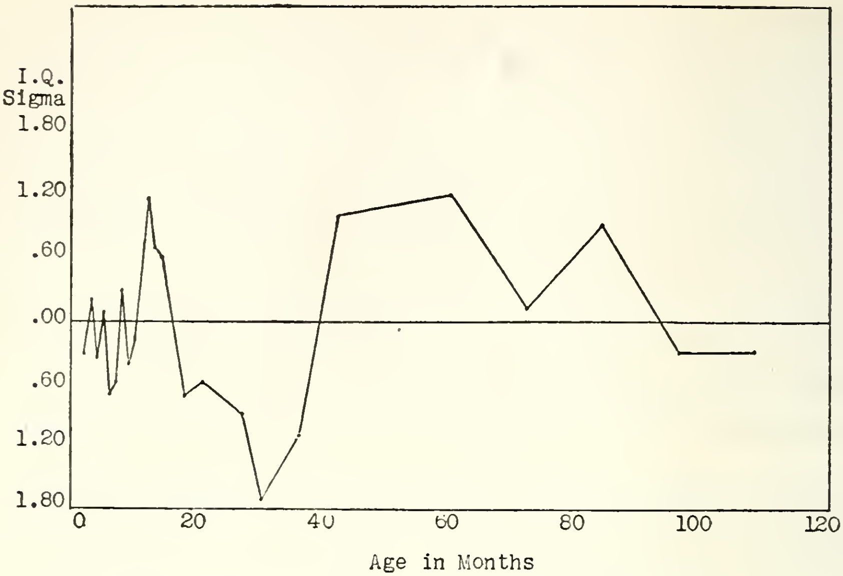
Bayley (1933) studied 49 infants from birth to three years, using items from scales by Kuhlmann (1922), Gesell (1925) and Jones (1926). Each infant was given at least six examinations. The result showed such low correlations between items and between trials that Bayley states, "The behavior growth of early months of infant development has little predictive relation to the later development of intelligence, — even though the later behavior may depend in large part on the previously matured, elementary neural connections or behavior patterns." Correlations between total scores on consecutive trials ranged from .71 to .89 with a mean at .82, but the correlation between age and test total during the first year was .98. These high correlations between consecutive trials were probably due to differences in age level, rather than to similarities in mental organization. If the ages of the infants were held constant, the correlations between trials would become nearly zero.

Figure 31A illustrates the changes in IQ reported by Bayley (1940) from successive tests of one child over a period of nine years. Bayley summarized a study of forty-eight children from 1 month to 9 years of age as follows:

> In an attempt to find some measure of mental ability that would rate children consistently during the first nine years, several combinations of scores were studied.
>
> (1) A "Developmental Score" made up of the sum of the items of the separate mental and motor tests yielded no greater consistency than did the mental test alone.
>
> (2) A selection of items from the California Preschool Mental Scale improved slightly the correlations of the two- and three-year tests with later performance, but the improvement was not great enough to produce constant scores over four or five years.
>
> (3) Tests of vocabulary given at from six to nine years of age gave scores moderately related to language tests at three and three and a half years, and were not significantly related to the age of first talking or to early mental-test scores.
>
> (4) Tests of formboard and puzzle-board performance (at five and a half to eight and a half years), although related to mental-test and vocabulary scores at similar ages, were unrelated to tests of ability during the first year.
>
> It was concluded that mental organization changes with growth, and that the rate of change is especially rapid before two years of age.

Bayley also reported that educational levels of the mid-parent, that is an average of both parents, and the scores of infants between 1 and 15 months of age are slightly negative. With further growth, however, the relationship changed and at 3 years of age reached a positive correlation of .47.

Furfey and Muehlenbeim (1932) used the Linfert-Hier-



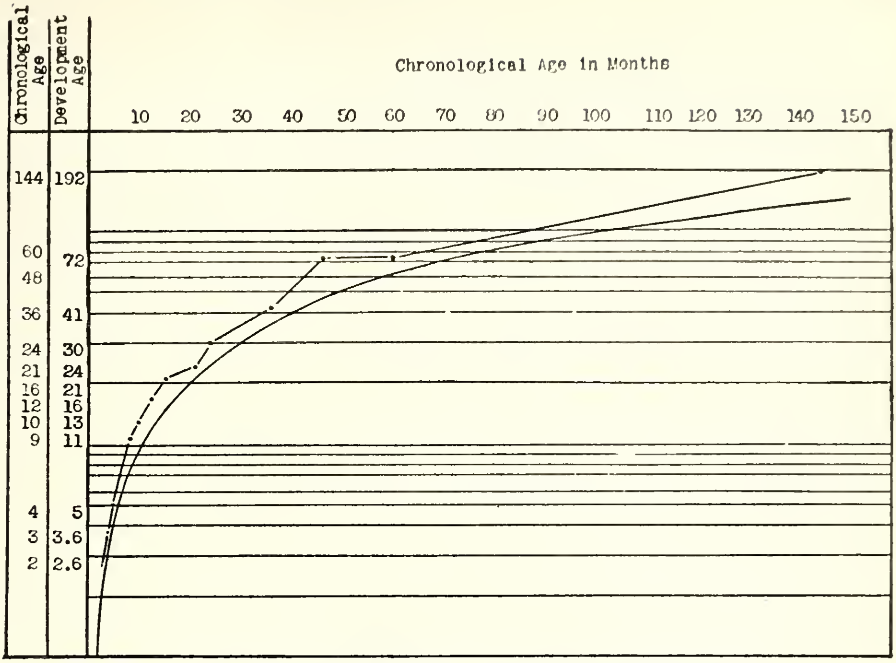
Note: This child varies from the average from +1.20 to -1.80 Standard Deviations by the age of 30 months. Variations of this degree were fairly common among preschool children.

FIGURE 31A. **Growth Curve of Early Childhood**

(Bayley, 1940, p. 25. By permission of the National Society for the Study of Education.)

holzer Scale (1928) with groups of infants who were 6, 9, and 12 months of age. Later eighty-one of these children were tested in the Stanford-Binet (1932) scale at the age of 4 years, 3 months. The correlation between the two scales was .00±.07. This zero correlation may mean that the two scales were measuring uncorrelated patterns. It may also be due to variations in individual growth or organization at various ages. Most of the tests for infants are heavily weighted with manipulative skills, whereas the Stanford-Binet requires much use of language at the age of 4 years.

Macfarlane (1938) concluded from a four-year study of 244 preschool children that mental ability, as indicated on the California preschool scales, showed less predictability as the interval between tests increased. The correlations between tests separated by six-month intervals ranged from



Note: This child maintains a position slightly above average with very little variation. The average is shown by a logarithmic curve used by Gesell.

FIGURE 31B. **Growth Curve of Early Childhood**

(Gesell, 1940, p. 152. By permission of the National Society for the Study of Education.)

.68 to .72. The correlations between tests given at 21 and 72 months were approximately .30. Macfarlane also found very low correlations between the child's mental test score and midparent's education. These ranged from .07 at 21 months, to .12 at 36 months, to .33 at 72 months. No simple explanation of these results is apparent. Since the test material included a wide variety of mental and motor skills, any simple interpretation is improbable. It may be that the content of the tests changed with age, or that the organization of traits in the child became more stable.

Gesell (1929, 1940) reported that a diagnosis of mental defect might safely be made by an expert worker during the first year of the subject's life. His reports are based on clinical studies and do not utilize correlation analyses of items or test scores.

Figure 31B shows one of Gesell's typical records. Here the developmental age shows very regular progress from 2 months to 12 years. He concludes:

> The foregoing cases were selected from some 10,000 now on file at the Yale Clinic of Child Development. Numerous biographic case studies by members of the staff have strengthened the conclusion that the basic trends and tempo of behavior development, as a rule, manifest themselves in infancy. If a child has normal growth potentialities, it is almost certain that they will reveal themselves to clinical perception in the first two years of life. Temporary "irregularities" of development are more frequently encountered throughout the preschool years, because of the nascency and interdependencies of behavior patterns during this formative period. In a few extremely exceptional instances, bound up with obscure emotional and physical factors, the signs of normality may be delayed as long as three years. On the other hand, virtually every case of primary feeblemindedness can be diagnosed in the first year of life. In a wide clinical experience we have never seen a case of secondary feeblemindedness due to educational or environmental deprivation, although we have seen an occasional case in which the I.Q. itself had descended to an apparently defective level.

Another study by Gesell (1937) reported work on five normal 1-year-old infants from superior homes. He found that fifteen behavior traits which were estimated on the basis of observations of films alone showed a marked correspondence with the results of a direct study four years later. He concludes that certain fundamental traits, which probably rest on biological characteristics, appear early and assert themselves consistently throughout this period.

It seems clear that all of these studies were based upon too few cases to be conclusive and upon samplings of patterns of performance which were scored and totaled by different methods. Nearly all the investigators point out that similar patterns are not tested at different ages. Hence comparisons from month to month or year to year are inconclusive. However, much careful work has been done and future investigations can exercise more adequate controls because of it.

9

MEASURES OF ACHIEVEMENT

I. CARDINAL OBJECTIVES

Statements of cardinal objectives in education have been offered by hundreds of authorities representing various creeds and political convictions. A clear statement of six democratic objectives, which may be considered typical for the whole country, was issued by a committee of superintendents, principals, and teachers in New York State in 1931. These six objectives are listed in Table 29, together with a list of instruments used by Wrightstone (1938) to appraise progress toward these goals in two different elementary school curricula. Some of these instruments of appraisal have been developed much further than others.

The instruments named in Table 29, Section C, *Critical Thinking* and, Section E, *Knowledge and Skills,* are well developed, and there are available several other equally effective tests which might be substituted for those listed. For a study at the high-school and college level the list would doubtless be extended to include tests of technical information and skills. These types of instruments are traditionally called Achievement Tests, and samples of them will be de-

TABLE 29

Cardinal Objectives of Elementary Education and Instruments of Appraisal

Cardinal Objectives of Elementary Education	*Instruments of Appraisal*
A. PRACTICE DESIRABLE SOCIAL RELATIONSHIPS	
1. Information about important elementary current affairs of social, economic, and aesthetic nature	1. Specially constructed information test entitled *Current Affairs*
2. Broad-minded social beliefs and attitudes related to topics and phenomena studied in the elementary school	2. Specially constructed test on social beliefs and attitudes, entitled *What Do You Believe?*
3. Honesty in scoring own tests and papers	3. Maller, J. B., *Self-Marking Test* (*of Honesty*)
4. Performance of acts indicative of initiative, co-operation, leadership, and work spirit.	4. Controlled-observation and anecdotal record techniques: See "Constructing an Observational Technique" by J. W. Wrightstone, *Teachers College Record*, Vol. 37, October, 1935, pp. 1–9
B. DISCOVER AND DEVELOP INDIVIDUAL APTITUDES	
1. Development of aptitudes, such as telling and writing stories, writing poetry, drawing and painting, construction, music, and dancing	1. (a) Specially constructed questionnaire; (b) *Judgment Rating Scale for Creative Writing;* (c) *Judgment Rating Scale for Creative Graphic Arts;* and (d) Meier-Seashore *Art Judgment Test*
C. CULTIVATE POWERS OF CRITICAL THINKING	
1. Finding and selecting data on problems	1. Specially constructed test, entitled *Obtaining Facts*, for social studies
2. Drawing conclusions and inferences from data	2. Specially constructed test, entitled *Explaining Facts*, for social studies
3. Applying conclusions or generalizations to events	3. Specially constructed test, entitled *Applying Facts*, for social studies
D. APPRECIATE WORTHWHILE ACTIVITIES	
1. Participation for enjoyment in such activities as art, music, reading, games, experimentation, and travel	1. Specially constructed questionnaire on participation in voluntary extra-school leisure activities
E. COMMAND OF COMMON INTEGRATING KNOWLEDGE AND SKILLS	
1. Reading silently with ease, speed, and comprehension	1. (a) *Metropolitan Reading Tests, Primary II:* Form A (b) *Stanford Reading Test:* Form W
2. Spelling with accuracy the normal vocabulary	2. (a) *Metropolitan Spelling Tests, Primary II:* Form A
3. Writing and speaking with freedom from gross errors	3. (a) *Metropolitan Language Tests, Primary II:* Form A (b) *Stanford Language Test:* Form W
4. To perform accurately in arithmetic the four fundamental operations and to know when to use them	4. (a) *Metropolitan Arithmetic Test, Primary II:* Form A (b) *Stanford Language Test:* Form W
F. SOUND PHYSICAL AND MENTAL HEALTH	
1. (a) Information about proper physical health habits and attitudes, first aid, and effect of harmful drugs and narcotics	1. (a) Specially constructed tests on *Health Awareness*
(b) Physical strength and fitness index	(b) Rogers, F. R., *Physical Capacity Tests*
2. Mental health habits and attitudes indicating (a) emotional stability and self-control and (b) friendliness or sociability	2. (a) Specially constructed *Personality Trait Indicator:* Part I — Personal Adjustment Part II — Social Adjustment

(Wrightstone, 1938, p. 163. By permission of the Bureau of Publications, Teachers College, Columbia University.)

scribed in this and the next chapter. The appraisals listed in Section A, *Desirable Social Practices,* Section D, *Appreciation of Worth While Activities,* and Section F, *Mental*

Health, are only moderately well developed. They are described in Chapters 15 to 19 under the titles of *Interest, Attitudes,* and *Modes of Behavior.* The instruments listed in Section B, *Aptitudes,* show various stages of development. Tests of intelligence which might logically be included here are described in Chapters 10 and 11; appraisals of mechanical ability, in Chapter 12; and appraisals of artistic ability, in Chapters 13 and 14.

Achievement tests at the earlier grades usually concentrate on basic language and arithmetic skills. At the junior high-school level special memory tests of school subjects are introduced. At the senior high-school and college levels much specialization is reflected in the large number of tests related to professional and occupational studies. Four main groups of special achievement tests will be described: physical studies, social studies, foreign languages, and business studies. A number of appraisals of study habits or methods of work will also be illustrated.

No attempt has been made here to list the most valuable tests in any field because as yet only a few fragmentary studies of the relative merits of various tests are available. One should consult yearly reviews of tests, such as Buros' *Mental Measurement Year Book* and the *Review of Educational Research,* for more detailed criticism and information. Illustrative material has been included here on the basis of wide usage, reliability, and interesting diagnostic possibilities.

II. MEASUREMENT OF LANGUAGE

Language, broadly defined, is any series of oral or motor acts by which individuals communicate with each other. Language may be divided into two general classes: unlearned signals for action, and symbols which refer to some experience. Animals have signal languages made up of movements and sounds which cause particular responses in other animals. These are often calls which lead to protection, flight, mating, the sharing of food, or other experiences. Such sounds seem to be developed through maturation along with the acts and without any intention on the part of the animal to give signals. For instance, animal psychologists usually agree that in a fight a dog barks or growls, not to scare an opponent, but because the situation sets off a series of movements which result in barking. Similarly, a mother hen does not cluck because she intends to tell the chicks that she has found food, but because the situation brings out a clucking response. Such explanations of communicative behavior among animals have been well supported by careful observations. The sounds made by small infants seem to be of this same unpremeditated variety, and among adults gestures and modes of expression often appear to be unlearned or only slightly modified by learning. Such elements of communication are important but so difficult to appraise that they are seldom recorded in test results.

A symbol is defined as an act or object which becomes a substitute for another. A symbol may or may not be intentionally given. Any symbol gets its meaning from the responses which it evokes. Sounds and words develop their common meanings by social agreement. Thus, the word *hot* is a symbol agreed upon by a particular group of people to refer to a common pattern of experience. Authorities usually agree that written language originated partly from pictures and partly from sound symbols. Some spoken words probably began as imitations of characteristic sounds and as unintentional signals.

Psychologically, the most significant aspect of language is probably not the form, although that is important, but the processes of abstraction and combination. The isolation of a particular aspect of a situation is called an *abstraction.* When an abstraction is experienced in several situations and remembered, it becomes a *concept.* The simplest concepts are experiences of contrast or similarity of size, shape, length, loudness, etc. Concepts may also be very intricate patterns which depend upon several senses and which combine other concepts. Complex concepts are illustrated, for instance, by the rules of tennis and by the meanings of the words *federalism* or *maturation*. Not all concepts are expressed or remembered verbally, but all words refer to concepts.

A complete list of important language factors would include the following somewhat independent skills and knowledges:

1. Techniques of expression: speech, grammar, punctuation, spelling, rhetoric, and penmanship
2. Word knowledge in various fields
3. Reading: complex skills of perception, comprehension, and reasoning
4. Muscular co-ordination, as in speech and writing
5. Visual acuity, as in reading and writing
6. Auditory acuity, as in speaking and listening
7. Knowledge of authors, publications, literary style, and form
8. Attitudes toward language usage and literature

The first three of these items cover the principal skills of language achievement tests; hence they are described here in some detail. The next three items, Nos. 4, 5, and 6, are not dealt with in this book, inasmuch as they involve very elaborate techniques for diagnosis and remedial treatment. The last two factors, which concern literary discrimination and appreciation, are discussed in Chapter 14.

A. Appraisals of Expression

Expression, either oral or written, although one of the most important and usual aspects of a person's behavior, is very hard to evaluate. Whereas aspects of expression, such as handwriting or pronunciation, can be mechanically recorded, the evaluation of these is always in the mind of the reader or listener. Moreover, some of the poorest expressions from a logical or grammatical point of view have been very popular and very effective politically, socially, and even artistically. However, widely accepted standards for speech, handwriting, and composition are available.

1. Speech. Two sorts of appraisals of speech have been found useful: standardized ratings of speech in public speaking situations, and tests for speech defects. Travis' (1935) diagnosis of speech needs in public speaking, Table 30, gives a fairly complete basis for a rating scale.

TABLE 30

Diagnosis of Speech Needs and Abilities

Diagnosis in speech has for its main purpose the determination of the individual's abilities and disabilities in speaking, in order that a program may be established to fit individual needs. In such a diagnosis the following factors should be taken into consideration:

1. *Choice of Topic.* Does the student choose topics, the immediate purpose of which is to inform the audience about something? Is the topic worth talking about, timely, of interest to the speaker and listener?
2. *Choice of Material.* Is the material adequate for the general purposes of the speech? Is it carefully analyzed and carefully selected for each point in the speech? Does the student speak from a background with which he is intimately familiar?
3. *Organization of Material.* Is there a purpose; a plan; a main idea; proper partitions and transitions; subordination of details; adaptation of thought and material to the purpose, audience, and occasion; and an effective conclusion?
4. *Symbolic Formulation and Expression.* Is the student able to utter a coherent series of ideas expressed in language that is adequate and in a form that is essentially correct in regard to grammar, sentence structure, and vocabulary?
5. *Projection to the Audience.* Is the student able so to stimulate the audience as to get adequate response? Is he able to secure and hold attention in a specific speaking situation?
6. *Overt Bodily Activity while Speaking.* Does the student have poise and co-ordination in bodily movement? Is his bodily activity purposive, relevant, and effective?
7. *Rhythm.* Does the student speak fluently and rhythmically, without interruption of the flow necessary for good speech? Is his speech free of repetitions, hesitancies, blocks, and stuttering?
8. *Articulation during Speech.* Are the speech sounds correctly produced?
9. *Voice and Voice Control.* Does the student's voice possess a basic quality that is clear, full, rich, resonant, and pleasing? Is the pitch of his voice low, medium, or high? Is his range narrow or broad? Is the pitch pattern flexible or rigid?
10. *General Effectiveness.* Is the student effective in terms of the total stimulus pattern set for him?

(Travis, 1935, p. 401. By permission of the National Society for the Study of Education.)

Standardized tests for speech defects include rough appraisals of oral reading or spontaneous speech and also very detailed tests of articulation. Gray (1917) and Stinchfield (1923) prepared tests of oral reading which require one to speak all the common English sounds. More detailed tests have been designed by Bryngelson (1932) and Durrell (1935). The careless child can be distinguished from one with a defect in speech by tests of ability to imitate speech sounds from the standard reading tests. Diagnosis and remedial treatment of serious speech defects require highly specialized techniques; hence they should not be undertaken without special training. Teachers and parents can aid a child in developing clear speech by setting good examples, avoiding emotional tensions, and praising good expression.

2. HANDWRITING. Three sorts of evaluations of handwriting are fairly common. In the first, experts try to determine for legal evidence whether or not two or more samples were written by the same person. Here the ways in which a person's handwriting may vary under different conditions are investigated. In the second, graphologists try to find indications of personality traits from samples of handwriting. In the third, the excellence of handwriting is rated according to scaled samples of penmanship.

The work of graphologists depends upon clues which, after long experience, may be interpreted as indicators of characteristic modes of behavior. Although such interpretation is often associated with charlatanism, still the best graphologists can reveal traits which are shown to be present by other methods. For instance, Cantril and Rand (1934) located six individuals who showed very high scores in one and only one of five parts of the Allport-Vernon Tests of Values: Aesthetic, Economic, Theoretical, Political, and Religious, see Chapter 16. Each of these persons then wrote the same letter on uniform paper and signed a fictitious name. When the six letters were photostated and submitted to 24 graphologists, 17 of them indicated four or more of the students' main interests correctly, an occurrence that would theoretically happen about once in a million times by pure chance. When the six letters were submitted to 26 educated adults who had no knowledge of graphology, no adult succeeded in naming the main interests of as many as four students correctly. In another experiment reported by Harvey (1934), samples of a prose selection were written by 50 college women; then twenty-six direct measures of script of each student were correlated with scores on Thurstone's Personality Schedule and Allport's Ascendance-Submission Tests. The relative heights and area of letters proved to be fairly consistent. Three graphological measures, *capital ratio, loopiness,* and *line overlap,* correlated significantly with both tests, and the multiple correlations of the three with the Thurstone Schedule was .89 and with the Allport A–S Test −.56. A very fascinating field of research lies here.

Scales of penmanship were among the earliest to be developed. That of Thorndike (1910), which seems to have been the first, consisted of samples reproduced with their scale values assigned by the equal-appearing-interval technique (see Chapter 23). Ayres (1912) used a similar method to devise a scale which has been very widely used because he published grade norms for both quality and rate. Figure 32 shows samples from his scale for three scale values, 20, 50, and 80. The averages for quality were found to increase from 38 at Grade II to 62 at Grade VIII, or four points for each grade. The rate in average words written per minute increased from 31 at Grade II to 79 at Grade VIII for the standard selection. Koos (1918) found the average quality of handwriting of adults in twenty-five different occupations to be 49.5 on the Ayres Scale, which is about the Grade V level.

Several check lists for a systematic recording of aspects of penmanship have appeared. Freeman (1914) devised one which not only listed defects but also noted their most common causes. Nystrom (1930) designed seven scales: for

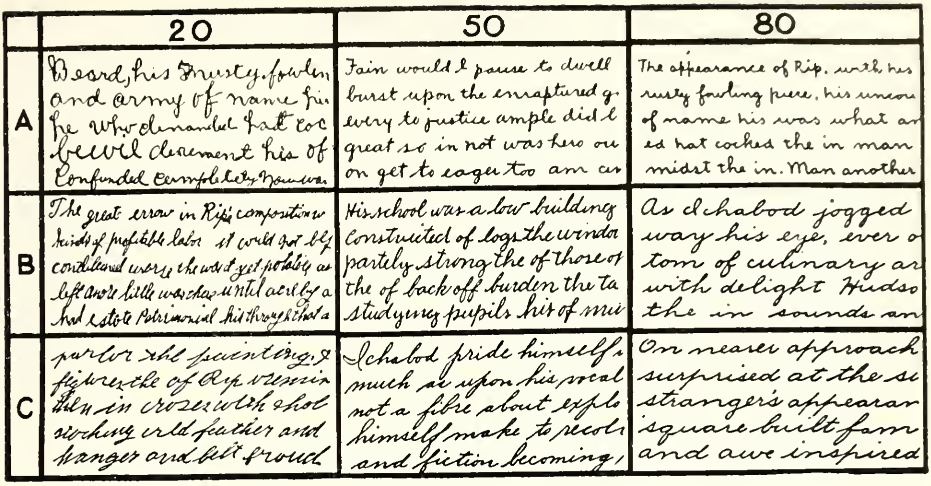


FIGURE 32. **Ayres Handwriting Scale**

About One-Third Actual Size

(Ayres, 1912. By permission of the Russell Sage Foundation.)

color or heaviness, size, slant, letter spacing, word spacing, beginning and ending of strokes, and alignment. Remedial suggestions and diagrams to aid students accompanied each scale.

3. Scales of English Composition. In order to provide a more reliable means of grading English composition than the judgment of a single teacher, Hillegas (1912) constructed a scale for use in Grades IV to XII. It consisted of a series of short essays arranged in order of excellence by a group of judges and numbered to show standard deviation scale values and grade norms. About a dozen similar scales have appeared, the most analytical of which is undoubtedly that of Van Wagenen (1923). He provided different scales for:

1. Exposition: 16 short essays on the topic "How I Earned Some Money"
2. Narration: 15 essays on "When Mother Was Away"
3. Description: 16 essays on "It Was a Sight Worth Seeing When the Troops Marched Away"

Each essay is rated for three qualities, *Thought Content, Structure,* and *Mechanics,* and the average rating is taken as an indication of general merit. In order to aid the rater, his attention is called to particular aspects of a composition, thus (p. 2):

> In rating for thought content in description, take into consideration:
>
> Maintenance of point of view (both physical and mental)
> Vividness of picture
> Emotional reaction
> Vigor and originality of diction
>
> In rating for sentence and paragraph structure, take into consideration:
>
> Unity
> Coherence
> Emphasis
> Variety and complexity of sentences
>
> The writer who uses many complex sentences shows a greater maturity of mind than the one who uses very simple or unnecessarily compound sentences, even though, from the very fact of the greater complexity, he may make more actual mistakes in structure.
>
> In rating for mechanical errors, take into consideration:
>
> Spelling
> Punctuation (only cases of actual error, not cases where punctuation is optional)
> Capital letters
> Paragraphing (only cases of actual error, not matters of preference)
>
> Here, too, one must take into consideration the range of vocabulary and complexity of expression. For instance, it is a more fundamental error to misspell "receive" than to misspell "psychological." Punctuation of very simple sentences would give less opportunity for error than that of complicated sentences or conversation.

Samples from the Description Scale, which runs from 0 to 100, are shown in Table 31. The scale values used by Van Wagenen were established by submitting the compositions to approximately 119 experienced teachers who arranged them in order of excellence three times; once for *Thought Content,* once for *Structure,* and once for *Mechanics.* A scale value of 10 was arbitrarily assigned to a

TABLE 31

English Composition Scales: Description Scale Items

General Merit 7.3

Thought Content 8 *Structure 9* *Mechanics 6*

It was a Sight worth seeing when the Troops marched past.
Went thet marched past is was fun to watch then. And a puch of solirden narched past.
And a puch of trump with flag and peolle with there flag.
And then cant a puch of boy and girls with cines of flag.
And there were twelte elenfeet.

General Merit 49

Thought Content 50 *Structure 49* *Mechanics 49*

It was a Sight Worth Seeing When the Troops Marched by.

It would send a thrill right through your body to see the troops march by. With the drums beating and the band playing it would make any body wish to join in with the kaki clothed men. All the soldiers looked like bunch of boys going to a sunday school picnic instead of going to the gloomy trenches. It is wonderful to see the soldiers keep time. The soldiers look as if they can't wait until they get over there.

General Merit 76

Thought Content 76 *Structure 79* *Mechanics 73*

It Was a Sight Worth Seeing When The Boys Marched By.

The boys were going. It was hard to believe. It was hard to realize ninety two boys from our own high school were marching before us for the last time before they went to France. Line after line, and rank after rank passed us as we stood looking in amazement.

The day was wonderful. There was not a cloud to be seen in the light blue sky. There was a soft breeze from the south. The prime of our Indian summer was here and, the boys were marching past.

Mothers, sweethearts, sisters, fathers, and brothers stood watching them as they went by. It was a sorrowful day for many of the onlookers, because probly some of them said good-bye for the last time and never to be welcomed home again. Still in the heart of each mother there was a little pride which lifted their heads a trifle higher.

An aged woman stood next to me and she said to me, "My but I hate to see Jimmy go. He's the only one I have left, but — but I'm sort of glad he is going, because its — its a wonderful thing hes going to fight for." Amidst her sobs she could say no more.

Many others stood crying. A flag waved from every rank, from everyery window, and from nearly every little child.

As dusk came on everything became quiite. The boys boarded the train, the people went home, everything was quite, the boys were gone.

(Arranged from Van Wagenen, 1923. By permission of the Educational Test Bureau, Minneapolis, Minn., and the World Book Co.)

difference between two items which caused 75 per cent of judges to rate one item better than the other. The whole scale extends over ten such differences. These samples have been chosen to show nearly the same scale values for *Thought Content, Structure,* and *Mechanics,* but in school situations a composition is often found to be high in one aspect and low in the others. This diagnosis allows a teacher to aid a pupil in the most effective manner.

Since letter writing constitutes the major part of the written expression of perhaps nine-tenths of adults, the scales for grading the general excellence of correspondence prepared by Lewis (1923) are of considerable interest. Five separate scales of nine items each have been assembled for (*a*) letters ordering material, (*b*) applications for a position, (*c*) narrative social letters, (*d*) expository social letters, and (*e*) simple narrations. He suggests that pupils be allowed to compare their own compositions with those of the scale, so that they may note the good points and correct their errors. Another special scale of considerable interest is that of Stewart (1934), who published graded samples of news stories from high-school journalism classes.

Netzer (1938) began the standardization of an oral composition scale for Grades IV, V, and VI after recording a large number of children's responses to pictures, incomplete stories, and objects. The children responded "best" to the objects, next to the stories, and third to the pictures. Much more work along this line is needed.

The task of grading essay-type examination questions is very similar to that of appraising English compositions. Since this type of examination is probably the most common of all types, methods of grading it have been widely studied. A good many reports have appeared showing that one examiner usually differs considerably in assigning grades to essay examination answers on two different occasions. Common self-reliability correlations, given by Stalnaker (1936), range from .40 to .60 in high school and college classes.

Correlations between separate examiners have often been found to be less than these figures owing to different standards of grading for facts, inferences, and grammar. On the other hand, single readers of the Regents' and of the College Entrance Board Examinations have usually shown self-reliabilities of .90 or more, and similar correlations between different readers are the rule. Stalnaker (1937) and Wrightstone (1937) have given detailed instructions for consistent grading. These include:

1. The question should require only one very definite and restricted answer, such as a statement of fact, or of an attitude, or an interpretation.
2. If several purposes or skills are to be graded, these should be graded one at a time.
3. The ideal answer should be carefully formulated and credits for partial answers agreed upon by all the judges.

Traxler and Anderson (1935) found that reliability of grading two English essays of high-school students was high when carefully done but that the retest-reliability of the pupils was relatively low over a short period of time. Much research is needed to show the usual relationships among such language skills as vocabulary, grammar, reasoning, and style.

B. Appraisals of Comprehension

The task of measuring comprehension is usually much easier and less controversial than that of measuring expression because a comprehension test can use a multiple-choice technique and because experts can usually agree upon the scoring. Literally hundreds of tests of English Comprehension have been developed which can be scored by a clerk without knowledge of the subject or even by a machine. These tests which are widely used will be described under the headings: *Word Knowledge, Techniques of Expression,* and *Reading Ability.*

1. Measurement of Word Knowledge. Knowledge of the meanings of single words is basic to all language skills; hence tests of vocabulary have become an essential part of all achievement tests. The only sure way to measure a person's vocabulary is to ask him to define all possible words. Since the unabridged dictionaries include nearly half a million English words, this would take a long time. It has been found, however, that fairly good examples of a person's vocabulary can be secured in a short time by the use of well-selected tests.

a. General Word Counts. In order to select important samples of words, a number of counts have been made of the words used most commonly in various communications. Thorndike (1927) and his staff produced an alphabetical list of 10,000 most commonly used words. These were found by tabulating about 4,000,000 words from newspapers, magazines, classics, novels, correspondence, and textbooks on common subjects. Each word was given an index number indicating its relative frequency. The *Thorndike Century Junior Dictionary* (1935), which lists and defines the 23,000 most common words, is an extension of this work to include a much larger sample of publications. A similar list of 10,000 words was produced by Horn (1926) based on the analysis of personal and business correspondence of adults, most of whom had received more than average education. Buckingham and Dolch (1936) have compiled a list of 19,000 words selected from eleven other lists and marked to indicate grade difficulty. Although the general agreement among various word counts is marked, different studies have yielded somewhat different relative frequencies of words. One of the reasons for discrepancies is the variation in words listed. Thorndike listed verbs and nouns from the same stem, as "contain" and "container," but not the modifiers, "contained" and "containing." He included as distinct units words which had the same root but somewhat different meanings and different frequencies, such as "constituency," "constituent," "constitute," "constitutional," "constitution," and "constitutionality." Persons who know one of such a series may be able to infer a correct meaning for most of the others. No thorough study has come to hand revealing the relationship between knowledge of roots and knowledge of words containing the roots. Likewise, since there is no standard and widely accepted method of counting words, it is not possible to speak of the size of a person's vocabulary in standard terms. The number of basic ideas needed for usual communication is much less than the number of words in Thorndike's list.

b. Special Word Counts. In order to examine vocabularies in special fields, word counts have been made of texts and other publications in both the physical and social sciences. An excellent summary of master lists of terms used in the social sciences was reported by Kelley and Krey (1934) and their associates. A list of 5,200 words is given which includes terms used in government, law, civics, political science, economics, religion, and sociology.

c. Minimum Vocabularies. Several interesting attempts to select basic vocabularies have been made in connection with the teaching of foreign languages and the writing of dictionaries. West (1935) found that an excellent English dictionary could be written using only 1,923 words, and that only 1,106 words were enough for an adequate speaking vocabulary. In a comparison with seven other minimum vocabularies, West found that all eight lists included only 2,219 different words. Some of the words which seemed important to West did not appear with great frequency in Thorndike's list, but in general there was a marked relationship between the frequency of words and their basic usefulness.

Minimum vocabularies have also been listed for special fields of information. Pressey's (1934) description of the selection of items essential for the understanding of history is typical of the best work. She first made frequency counts of words in six widely used history texts. These combined with the published results of others resulted in a master list of 1,444 words. This master list was then presented to sixty-nine teachers of history in secondary schools and colleges with instructions to mark each word *essential, accessory,* or *unimportant.* The entire list was next rated by seven individuals especially trained in social studies who indicated their judgment of the values of each word outside of the history classroom. Finally, a list of 415 words was selected all of which were frequently used and highly rated for both historical and sociological usage. Table 32 shows the words which were finally selected.

In an earlier work Pressey (1924) prepared lists of basic concepts in fourteen fields from the ratings of teachers concerning technical importance. The fields, which are not mutually exclusive, are:

1. Grammar and Composition: English, French, Latin, and German
2. Literature
3. Arithmetic
4. Algebra and Geometry
5. History
6. General Science
7. Biology
8. Chemistry
9. Physics
10. Physiology
11. Home Economics
12. Manual Training
13. Art
14. Music

The selection of essential terms at the high-school and college level for many different fields has now been completed by groups of interested teachers.

d. Selection of Words for Tests. A fairly large number of tests have been constructed to measure either general or special vocabularies, using the following procedure. A random sample of about 100 words is selected from a word list, and test items are constructed about these words. The items are tried out on various age groups, and the order of difficulty is determined by noting the percentage of persons who succeed in each item. A final selection is made to secure a wide range of items and enough items at each level of difficulty to give a fairly precise and consistent discrimination among all persons tested. One of the first widely used tests was made by Terman (1916). He began by selecting a word on every tenth page of an 18,000 word dictionary. After preliminary trials he selected two lists of 50 words each, to measure general vocabulary from the mental ages of 8 to 19 years. Terman and Merrill (1937) selected from these two lists 45 words which seemed adequate to appraise general vocabulary from the mental ages of 6 to 22 years. This seems the more remarkable in the light of the fact that a minimum of 30 words is needed to distinguish between all of these age levels. Usually only two words are needed to indicate one year's growth in mental age. The reliability of such vocabulary tests is generally very high, .90 or more. Vocabulary tests are typically the most consistent of all tests since they are not very much affected by speed, practice, or adjustment to the test situation. Page 305 gives the United States Army Alpha Vocabulary Test of Opposites.

Thorndike (1926) used his word frequency lists in devising a test of 110 items in which there were eleven levels of ten items each. The levels were scaled to represent equivalent steps, and the items in each level were chosen to be of nearly the same difficulty.

Although the vocabulary tests of Terman and Thorndike were intended for use in appraising general intelligence, they are in effect achievement tests, and tests very similar to these have been included in nearly every appraisal of educational achievement. More than forty of these have been standardized on a national scale. As a rule the general vocabulary tests have been found to correlate as high as .70 or more with total scores on achievement tests and intelligence tests.

If one examines the words included in many general vocabulary tests, he usually finds very few terms from scientific or artistic fields. At the easier levels, a large number of terms describing common objects and personal relations are found, and at the harder adult levels, literary, social, and business

TABLE 32

Words Selected as the Essential Core of Vocabulary in History

A. GOVERNMENTAL TERMS:

ambassador
authorities
consul
governor
king
minister
official
prime-minister
premier
police
president
representative
secretary
senator
sovereign
statesman
vice-president

assembly
bureau
board
cabinet
commission
committee
conference
congress
council
department
House of Representatives
league
legislature
parliament
senate
session

anarchy
commonwealth
communism
confederacy
despotism

democracy
empire
federal
government
imperialism
monarchy
republic
self-government
tyranny
union

city
colony
country
county
dominion
nation
province
state
territory
town

alliance
arbitration
diplomacy
foreign
international
negotiation
neutrality
pact
peace
powers
reciprocity
treaty

amendment
article
charter
constitution

act

bill
declaration
decree
document
law
legislation
measure
petition
proclamation
proposal
provision
report
resolution
restriction
statute

abolish
abdicate
adjourn
annex
appoint
authorize
compromise
concede
conciliate
confiscate
enact
enforce
grant
impeach
inaugurate
nullify
ratify
repeal
repudiate
sanction
veto

executive
legislative
judiciary

appropriation
appointment
budget
coinage
currency
customs
debt
duty
expenditures
greenback
mint
protective
revenue
tariff
tax
treasury

doctrine
issue
policy
reservation

centralization
civil
civil service
domestic
internal
interstate
local
municipal
states rights

administration
regime

capitol
patriotism

prohibition
reconstruction
referendum

B. POLITICAL TERMS:

campaign
candidate
caucus
convention
deadlock
delegate
nominate
opponent
plank
platform
politics
ticket

anti-slavery
abolitionist
democrat
federalist
political party
progressive
republican
socialist
whig

majority
minority
unanimous

conservative
partisan
radical

lobbying
patronage
spoils system

ballot
election
polls
primary
suffrage
vote

C. ECONOMIC TERMS:

business
commerce
commodity
company
competition
consumer
exploit
export
factory
goods
import
industry

manufacture
merchandise
production
property
raw material
rebate
shipping
trade

corporation
monopoly
trust

employee
employer
labor
strike
union

bank
bankrupt
bond
capital
credit
crisis
depreciation
finance

inflation
investment
market
panic
speculation
stocks

communication
public utilities
transportation

prosperity
wealth

TABLE 32—*Continued*

Words Selected as the Essential Core of Vocabulary in History

D. Sociological Terms:

aristocrat
peasant
slave
society

community
homestead
pioneer
plantation
settlement

rural
urban

census
inhabitants
population

negro
race

mob
riot

emigration
expansion
immigration
migration

emancipation
freedom
independence
liberty
oppression

education
institution
invention
reform

people
private
public
public opinion
standard of living

E. Legal Terms:

arbitrary
illegal
invalid
justice
legal
rights
unconstitutional

alien
citizen
exile

nationality
native
naturalization

appeal
case
convict
crime
decision
execution
injunction
judge

jury
testimony
verdict
violation
witness

court
jurisdiction
supreme court

bribery
corruption

fraud
graft

conspiracy
insurrection
rebellion
revolt
revolution
secession
sedition
smuggling
treason

F. Military Terms:

allies
belligerents
enemy
hostile
pirate

army
commander
confederate
cruiser
general
fleet
forces
naval

navy
officer
reinforcements
service
troops
submarine

marine
militia
recruit
soldier
veteran
volunteer

draft
mobilization

aggression
attack
battle
blockade
bombardment
campaign
contraband
defensive
embargo
evacuation
fortification

invasion
massacre
military
munitions
occupation
offensive
siege
strategic
surrender
victory
war

armistice
disarmament
indemnity
reparations

G. Geographical Terms:

boundary
continent
continental
coast
frontier
prairie

district
region
section

discover
expedition

exploration
navigation
voyage

conservation
natural resources

agriculture
irrigation
reclamation

Pan-American

H. Religious Terms:

clergy
missionary

creed

denomination
heresy
intolerance
persecution
tolerance

catholicism
protestantism

crusade

papacy
pope

I. Terms Referring to Chronology and Records:

ancient
century
current
decade
medieval

era
event
modern
period

history
records

propaganda
publicity

civilized
primitive

movement
precedent
tradition

(Pressey, 1934, p. 186. By permission of Charles Scribner's Sons.)

terms. This selection doubtless reflects the frequency of word usage in common communications and also the occurrence of words in dictionaries. Many examiners have, however, believed that general vocabulary tests were more favorable to those with academic interests than to those with mechanical, agricultural, artistic, or scientific interests. This belief has doubtless been a force in the creation of a number of special vocabulary tests. Teachers in special fields have also desired such tests.

The procedure for constructing tests in special fields is similar to that used for designing general vocabulary tests. A special master list is first obtained. Then the selection of words is made in such a way that an adequate sample of each level of frequency is thought to be included. The adequacy of the sample depends upon the use to be made of the test; hence it cannot be definitely known in advance. After a test has been applied, the adequacy of the whole test, or of each separate item, in discriminating between individuals may be ascertained by the methods outlined in Chapters 5 and 6.

The Iowa Silent Reading Test, by Greene and Jorgensen (1927) includes four separate vocabulary tests:

1. Social science (20 items)
2. Physical science (15 items)
3. Mathematics (15 items)
4. English literature and grammar (20 items)

The Progressive Achievement Tests, Intermediate and Advanced, Tiegs and Clark (1934), include separate vocabularies for the same four fields, each of which consists of 25 items. Total vocabulary scores are secured by adding the four subtest scores, see Table 36A.

The Michigan Vocabulary Profile Tests, Greene (1939), include eight divisions of terms:

1. Human Relations: mental and social processes and situations
2. Commerce: business, manufacture, sales, economics
3. Government: legislative, executive, judicial
4. Physical Sciences: physics, chemistry, mechanics

5. Biological Sciences: zoology, anatomy, pathology
6. Mathematics: arithmetic, algebra, geometry, trigonometry
7. Fine Arts: plastic, graphic, architecture
8. Sports: ten most common sports which adults play

Each division of the battery consists of 30 different items arranged in ten levels of difficulty. The items range in difficulty from those which were passed by at least 98 per cent of a group of college sophomores to those which were passed by 2 per cent or less of the same group. The gradation in difficulty of the items is such as to yield a test of high discriminative capacity at the senior high-school and college level.

Each item consists of a definition and four words or phrases, only one of which is completely and accurately defined or described. The subject is asked to select the one which he thinks is the correct one. Items which correlated less than .30 with total scores in their division were eliminated.

An attempt was made to eliminate all items in which the right answer could be found solely by reasoning from knowledge of roots and prefixes or by eliminating wrong answers. This attempt was not entirely successful, but the number of items of this sort has been reduced by using the same prefixes and roots more than once in an item and by selecting wrong answers which were nearly, but not quite, synonymous with the right answer. It was desired to make an information test which would be affected as little as possible by reasoning processes.

Reliabilities as indicated by correlations between two equivalent forms of 30-item tests given one week apart are shown in Table 33, in italics. These range from .78 to .94, with a median of .80. These figures show that longer tests are not necessary for fairly accurate individual predictions. If more accurate predictions are needed for some specific purpose, both forms might be used.

Table 33 also shows the intercorrelations among the various divisions of the Vocabulary Profile Test for a group of

TABLE 33

Intercorrelations of Divisions of Michigan Vocabulary Profile

Form A

Division	*Tests Designated by Number*								*Average*	*With Total* †
	1	2	3	4	5	6	7	8		
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
1. Human Relations	*.78* *	.24	.32	.16	.25	.35	.17	.14	.23	.61
2. Commerce		*.79*	.40	.51	.40	.42	.07	.27	.33	.57
3. Government			*.82*	.31	.37	.24	.05	.24	.27	.48
4. Physical Sciences				*.94*	.54	.50	.00	.49	.34	.76
5. Biological Sciences					*.91*	.10	.36	.18	.31	.68
6. Mathematics						*.86*	.23	.21	.28	.55
7. Fine Arts							*.79*	.10	.15	.44
8. Sports								*.80*	.23	.48

* Numbers in italic refer to correlations between tests of thirty items each, which were intended to be equivalent samples of a Division.

† Correlations of each Division with the total are a little too high because the totals contain all Divisions.

(Greene, 1939. By permission of the World Book Co.)

literary college sophomores. Results for other grade groups were substantially the same. These correlations are all below .55, with a median correlation of .27. The divisions thus show a large degree of independence, and they would appear even more independent in a large unselected population. Practically zero correlations are found between scores in Fine Arts and scores in Commerce, Government, and Physical Sciences. Scores in Physical Sciences correlated approximately .50 with scores in Biological Sciences and Mathematics. These figures indicate the presence of a number of fairly well-isolated factors. Psychologically there is little evidence for any functional relationship between the information in any two of these divisions, with the exception of Mathematics which is needed as a tool subject in many fields of human thinking.

Figure 33 shows graphically the score of Henry Brown on the Michigan Vocabulary Profile Test. There it appears that he was above the second-year college average in all except two fields: Fine Arts and Sports. His highest scores were in Physical Sciences, where he exceeded the scores made by the lowest 98 per cent of the group. He was also in the high 10 per cent of the group in Biological Sciences, Mathematics, and Total Test scores. Such profiles are valuable for appraising the technical information which a person has. They also predict fairly well his reading and composition skills in various fields.

Total Michigan Vocabulary Scores correlated .56 with the vocabulary section of the Co-operative English Test and .61 with the vocabulary section of the American Council on Education Psychological Examination for College Freshmen. Since all these tests have high self-correlations, the conclusion must be reached that they are measuring different and somewhat unrelated fields of information.

2. Techniques of Reading. In this division many well-prepared tests of spelling, grammar, punctuation, and sentence structure are to be found. Nearly all of these are

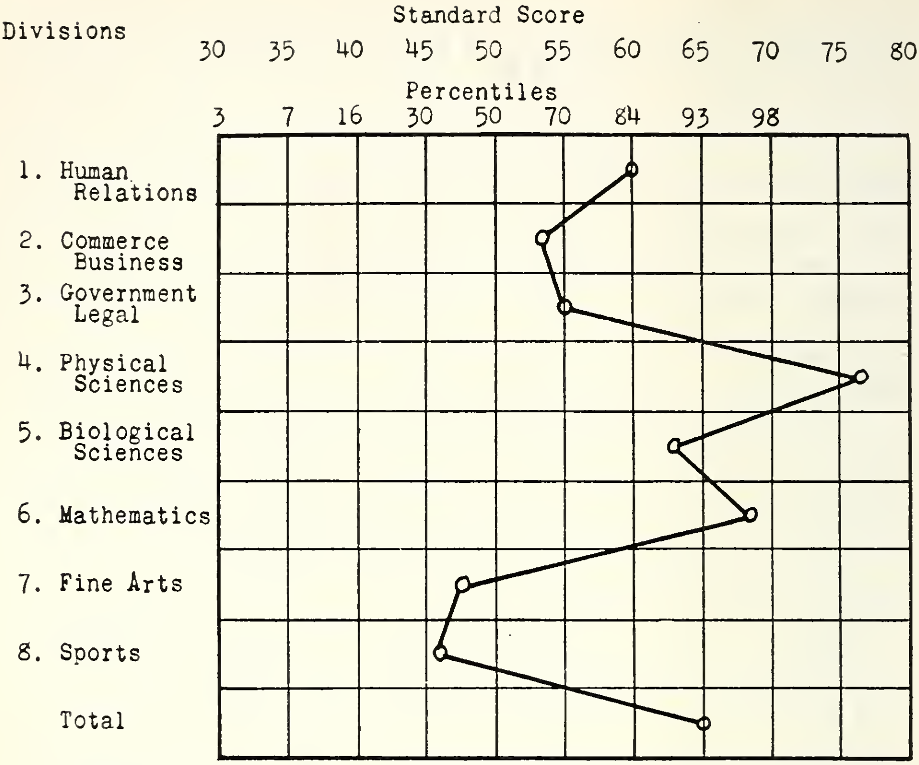


FIGURE 33. **Michigan Vocabulary Profile Scores of Henry Brown**

Norms for Second-Year College

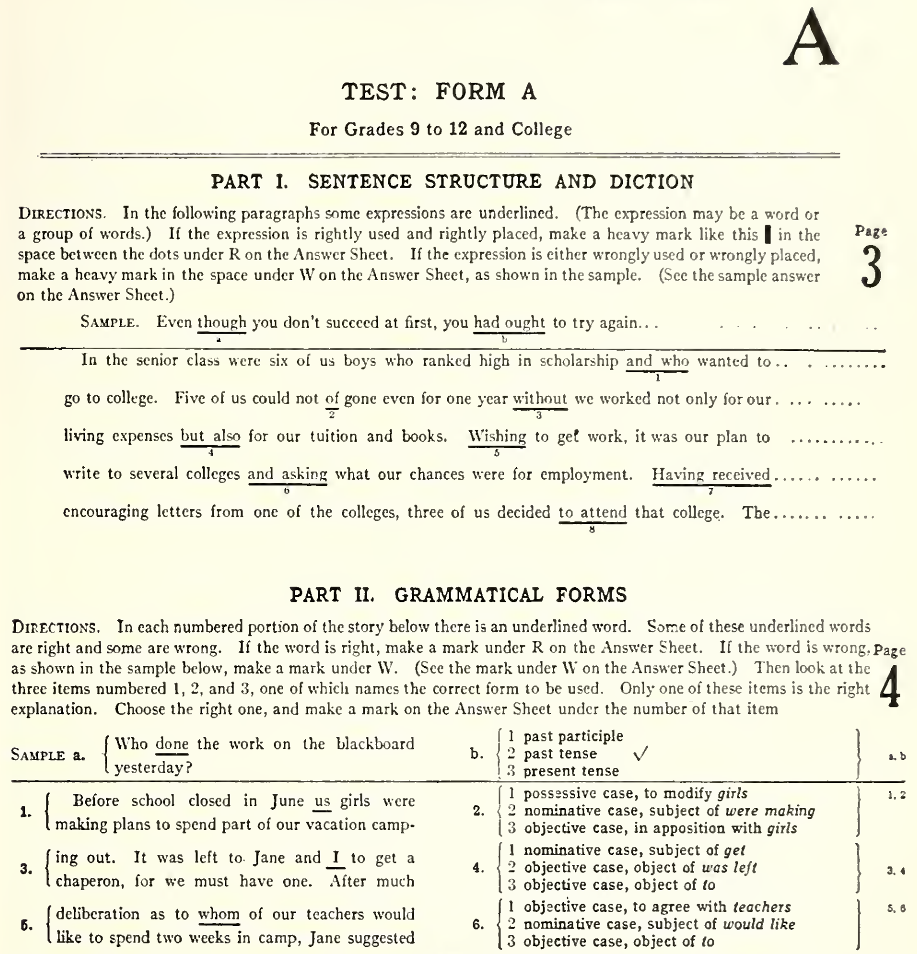
(Greene, 1939)

composed of items which contain errors. The student is asked to detect the errors and in some cases to correct them.

a. English Usage. The selection of items for tests of English usage has usually followed an investigation to reveal prevalent errors. The pioneer work of Charters (1920) shifted the emphasis from the rough scoring of papers to systematic diagnosis and remedial work. His studies showed that misuse of 14 common verbs gave rise to nearly 60 per cent of all oral errors. These errors were made by only 45 per cent of the students; hence individual rather than group treatment was suggested.

TABLE 34

Barrett-Ryan-Schrammel English Test



A

TEST: FORM A

For Grades 9 to 12 and College

PART I. SENTENCE STRUCTURE AND DICTION

DIRECTIONS. In the following paragraphs some expressions are underlined. (The expression may be a word or a group of words.) If the expression is rightly used and rightly placed, make a heavy mark like this ▌ in the space between the dots under R on the Answer Sheet. If the expression is either wrongly used or wrongly placed, make a heavy mark in the space under W on the Answer Sheet, as shown in the sample. (See the sample answer on the Answer Sheet.)

Page
3

SAMPLE. Even $\underset{a}{\underline{\text{though}}}$ you don't succeed at first, you $\underset{b}{\underline{\text{had ought}}}$ to try again. . . . .

In the senior class were six of us boys who ranked high in scholarship $\underset{1}{\underline{\text{and who}}}$ wanted to . . . .

go to college. Five of us could not $\underset{2}{\underline{\text{of}}}$ gone even for one year $\underset{3}{\underline{\text{without}}}$ we worked not only for our . . . .

living expenses $\underset{4}{\underline{\text{but also}}}$ for our tuition and books. $\underset{5}{\underline{\text{Wishing}}}$ to get work, it was our plan to . . . .

write to several colleges $\underset{6}{\underline{\text{and asking}}}$ what our chances were for employment. $\underset{7}{\underline{\text{Having received}}}$ . . . .

encouraging letters from one of the colleges, three of us decided $\underset{8}{\underline{\text{to attend}}}$ that college. The . . . .

PART II. GRAMMATICAL FORMS

DIRECTIONS. In each numbered portion of the story below there is an underlined word. Some of these underlined words are right and some are wrong. If the word is right, make a mark under R on the Answer Sheet. If the word is wrong, as shown in the sample below, make a mark under W. (See the mark under W on the Answer Sheet.) Then look at the three items numbered 1, 2, and 3, one of which names the correct form to be used. Only one of these items is the right explanation. Choose the right one, and make a mark on the Answer Sheet under the number of that item

Page
4

SAMPLE a.	Who $\underline{\text{done}}$ the work on the blackboard yesterday?	b.	1 past participle 2 past tense ✓ 3 present tense	a, b
1.	Before school closed in June $\underline{\text{us}}$ girls were making plans to spend part of our vacation camp-	2.	1 possessive case, to modify *girls* 2 nominative case, subject of *were making* 3 objective case, in apposition with *girls*	1, 2
3.	ing out. It was left to Jane and $\underline{\text{I}}$ to get a chaperon, for we must have one. After much	4.	1 nominative case, subject of *get* 2 objective case, object of *was left* 3 objective case, object of *to*	3, 4
5.	deliberation as to $\underline{\text{whom}}$ of our teachers would like to spend two weeks in camp, Jane suggested	6.	1 objective case, to agree with *teachers* 2 nominative case, subject of *would like* 3 objective case, object of *to*	5, 6

(Prepared by E. R. Barrett, Teresa M. Ryan, and H. E. Schrammel, Kansas State Teachers College, Emporia, Kansas. Copyright, 1938. By permission of the World Book Company.)

O'Rourke (1934), from a nation-wide study of English usage in Grades VII to XII, found that good sentence structure was the most difficult problem at all levels, that careless omissions or repetitions came second, and that ambiguous meaning came third. Errors in verb forms were uncommon, contrary to the reports of errors in the lower grades.

Table 34 shows portions of a typical test of language usage by Barrett *et al.* (1938) which yields scores for:

Part I. Sentence Structure and Diction (30 items, 10 minutes)
Part II. Grammatical Errors (35 items, 20 minutes)
Part III. Punctuation (30 items, 10 minutes)

Scoring is simplified by having the student place all his marks on an answer sheet, which may be checked very quickly. In this test, expressions in the printed text are underlined. The task is to determine whether or not the underlined expressions are rightly used. In Part II the student is asked to describe the correct form when a wrong grammatical form has been used. The test allows fairly liberal periods of work. Centile ranks are furnished which show means of 88, 96, 103, 106, and 109 for Grades IX, X, XI, XII, and college freshmen, respectively. Overlapping between Grade IX and college freshmen, which is very marked, is typical of reports of this sort.

Table 35 shows a proofreading type of test by Wilson (1923) in which 28 grammatical errors are to be corrected. Median grade scores rise from 7 at Grade III, to 22 at Grade VIII, and to 26 at Grade XII.

The question is often asked whether tests of the proofreading variety measure one's skill as well as scores taken from an actual composition. Willing (1926) made a comparison between a proofreading test containing 180 errors and an original composition. The composition revealed 12 errors per pupil, and the test, 59 errors. Although there were some errors in the composition which were not represented in the test, it seems clear that the test was a more comprehensive instrument. Both sorts of appraisals should be used for remedial work, and considerable research is needed to find usual relationships between them.

A few diagnostic tests have been prepared from which a detailed check of the specific types of errors made by a student can be secured. Leonard (1926) prepared a minute examination of punctuation and capitalization. Tiegs and Clark (1934) published the Progressive Achievement Tests Series, which contains the analysis of language errors shown

TABLE 35

Wilson Language Error Test

Directions: This is a test to see if you can correct the mistakes that a pupil has made in writing a short story. (A short sample is then given and corrected.) That is the way you are to do in this test. There are three stories in this folder. but you are to take only the first one unless your teacher tells you differently. You are to correct all the mistakes in that story just as it has been done in the sample. Draw a line through each wrong word and write the correct word above it. Be very careful to correct every mistake. Work at your usual rate. You will be given time enough to finish unless you are very slow. Be a good sport; do your best and play fair.

STORY A

Errors　　　　　　　　　　　　　　　　Corrections

Saturday Morning

✓✓ Saturday morning is a busy time ~~to~~ at are house. A feller has a good chance to work. ✓
Dorothy and I ~~Me and Dorothy~~ divide the tasks between us. Then we race to see who will finish ✓✓
✓ first. Last Saturday I ~~taken~~ took the breakfast dishes as one of my tasks. I am especial ✓
✓ fond of washing dishes. You should have saw me work. I wanted to get through
✓ so as I could play.
✓ John he called up at eleven o'clock to see if I might play with him. I had ~~too~~ two ✓
✓ rooms to dust before I could go. John saw that I couldn't hardly leave my work
until I had ~~did~~ done all of it. He brought over some doughnuts and gave them to me. ✓
✓✓ I sure appreciated the doughnuts. Then John helped me. It was real good of
him. When we had finished, I suggested playing marbles until time for dinner.
✓✓ "I ~~ain't~~ haven't got no marbles," said John. "They ~~comes~~ come very handy," I replied. Then ✓✓
✓ I ~~give~~ gave him some of mine. I had to many for my bag. John and I enjoy marbles. ✓
✓ When dinner was ready, mother invited John to stay. "If I was sure my mother
wouldn't care, I should like to stay," he replied. John ~~seen~~ Saw that he was really wanted ✓
so he telephoned to his mother. He enjoyed the dinner and ~~et~~ ate heartily. When ✓
✓✓ ~~them~~ the apples was passed, John wanted one, but he couldn't eat no more. After dinner ✓
✓ we had another game of marbles. I hopes John may come over again.

16　　　　　　　　　　　　　　　　12

(Wilson, 1923. By permission of the World Book Co.)

in Table 36 B in the lower part of the page. Seven situations in which errors in *Capitalization* may occur are listed for remedial action; six, in *Punctuation;* five, in *Words and Sentences;* and eleven, in *Grammar.* Although there are not enough items in each one of these subdivisions to yield reliable indications of special types of errors, the technique of diagnosis is probably sufficient for many situations. In

TABLE 36A

Progressive Achievement Tests — Advanced Battery Form A High School and College

(Diagnostic Tests keyed to the Curriculum)

Devised by Ernest W. Tiegs, Dean, University College, University of Southern California, and Willis W. Clark, Director of Administrative Research, Los Angeles County Schools.

Name James Brown Grade 10

School U. High Age 15-4 Birthday

Teacher Miss J. Date 12-12-38 Sex [✓ M] F

TEST SUBJECT	Possible Score	Pupil's Score	Percentile Rank for Grade...
1. Reading Vocabulary	100	46	50
A. Mathematics	25	12	40
B. Science	25	10	25
C. Social Science	25	14	60
D. Literature	25	10	55
2. Reading Comprehension	55	32	40
E. Following Directions	10	5	40
F. Organization	15	6	10
G. Interpretations	30	21	50
3. Mathematical Reasoning	60	38	80
A. Number Concept	20	12	70
B. Symbols and Rules	15	9	60
C. Numbers and Equations	10	8	70
D. Problems	15	9	85
4. Math. Fundamentals	80	48	40
E. Addition	20	15	55
F. Subtraction	20	14	40
G. Multiplication	20	9	35
H. Division	20	10	45
5. Language	125	74	30
A. Capitalization	15	12	40
B. Punctuation	10	6	70
C. Words and Sentences	25	21	45
D. Grammar	30	24	90
E. Spelling	30	5	5
F. Handwriting	15	6	20
TOTAL	420	238	45

DIAGNOSTIC PROFILE
(Chart Pupil's Percentile Rank Here)

Percentile Rank for Grade 10

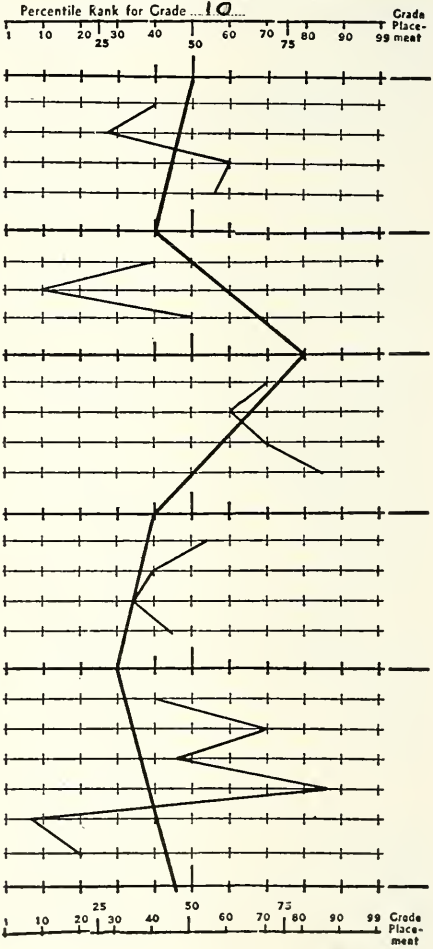


TABLE 36B

Diagnostic Analysis of Learning Difficulties

If the diagnostic profile on the first page of this test indicates that the pupil is making normal progress in all fields, the teacher will have no use for the following diagnostic analysis. However, where the diagnostic profile shows achievement below a desirable standard in one or more major fields, the following device will assist in isolating and analyzing the specific causes of difficulty as a basis for remedial instruction.

The numerals and capital letters in the diagnostic analysis correspond to the sections of the test similarly marked. For example, if the diagnostic profile shows unsatisfactory achievement in Test 4, Sec. E (addition in arithmetic fundamentals) an inspection of the unsatisfactory responses in this section of the test (by number) will reveal whether or not remedial instruction is needed in carrying, use of zeros, reducing to common denominators, and the like. These topics are then checked by the teacher as the basis for remedial work.

Once an adequate diagnosis has been made, remedial instruction is frequently a simple matter. However, teachers have in the past found the clerical work incident to following each individual pupil a heavy burden. Such extra work is almost completely eliminated if this diagnostic analysis is torn from the test booklet and kept on the teacher's desk, where the various items may be checked off as the pupil masters them.

READING

1. Reading Vocabulary

A. MATHEMATICS:
______ Basic vocabulary ...... ⑫ ...1-25

B. SCIENCE:
______ Basic vocabulary ...... ⑩ ...1-25

C. SOCIAL SCIENCE:
______ Basic vocabulary ...... ⑭ ...1-25

D. LITERATURE:
______ Basic vocabulary ...... ⑩ ...1-25

2. Reading Comprehension

E. FOLLOWING SPECIFIC DIRECTIONS:
______ Directions in mathematical situations ..1, 2, 5, 9, 10
______ Reading definitions and following directions..3, 4, 6, 7, 8

F. ORGANIZATION:
______ Vocabulary 1-6
______ Use of index............ 7-9
______ Selecting references 10-13
______ Report outline 14-15

G. INTERPRETATION OF MEANINGS:
______ Selecting topic or central idea 1, 10
______ Understanding directly stated facts......4, 5, 7, 8, 11, 12, 13, 14, 18, 22, 26, 28, 29
______ Making inferences......2, 3, 6, 9, 15, 16, 17, 19, 20, 21, 23, 24, 25, 27, 30.

MATHEMATICS

3. Mathematical Reasoning

A. NUMBER CONCEPT:
______ Writing integers 1-3
______ Writing money 4
______ Writing fractions 5-7
______ Roman numbers 8-10
______ Fractions and demicals.... 11-13
______ Exponents and roots........ 14-16
______ Negative numbers 17
______ Abstract numbers 18-20

B. SYMBOLS AND RULES:
______ Symbols 1-3, 8-10
______ Vocabulary 4-7
______ Rules 11-15

C. NUMBERS AND EQUATIONS.
______ Negative numbers............ 1-4
______ Solving equations............ 5-10

D. PROBLEMS:
______ Simple problems 1-2
______ Sharing and averaging.......... 3-4
______ Square and cubic measure 5-6
______ Insurance and discount.... 13-15
______ Ratio and percentage........ 7-11
______ Budgeting 12

4. Mathematical Fundamentals

E. ADDITION:
______ Simple combinations 1
______ Carrying 2-4
______ Zeros 1, 6
______ Column addition 3, 4
______ Adding money 4, 6
______ Denominate numbers 4-6
______ Reducing fractions to common denom. 8, 10-13
______ Adding mixed nos............ 10-13
______ Adding fractions and decimals 14-15
______ Writing decimals in column 16-17
______ Adding percentages 18
______ Adding abstract nos........ 19-20

F. SUBTRACTION:
______ Simple combinations 1
______ Borrowing 2-5
______ Zeros 1, 3, 5
______ Subtracting money 4, 5
______ Denominate numbers 4-6
______ Subtracting numerators........ 7-8
______ Reducing fractions to common denominators 9-10
______ Borrowing with mixed numbers 12, 13
______ Subtracting fractions from decimals 14, 15
______ Writing decimals in column 16, 17
______ Subt. abstract nos.......... 19, 20

G. MULTIPLICATION:
______ Tables 1-5
______ Zeros in multiplicand........ 2, 5
______ Zeros in multiplier........ 4, 5
______ Two-place multipliers 3-5
______ Cancellation of fractions 7, 9, 10, 11, 13
______ Fractions and mixed numbers 12
______ Fractions and decimals........ 15
______ Pointing off decimals...... 16, 17
______ Mult. abstract nos.......... 19, 20

H. DIVISION:
______ Tables 1-5
______ Zeros in quotient............ 1-4
______ Remainders 5
______ Inverting divisor in fractions 6-13
______ Mixed numbers 11-13
______ Reducing fractions to decimals 14
______ Pointing off decimals........ 15-17
______ Div. abstract nos.......... 19-20

LANGUAGE

5. Language

A. CAPITALIZATION:
______ First word of sentence............ 1
______ Names of persons............ 2, 7, 9
______ Names of places.......... 2, 3, 8, 9
______ Days of week and months.. 4, 6
______ Titles 5, 7
______ First word of quotation.......... 6
______ Over capitalization X

B. PUNCTUATION:
______ Commas X
______ Question marks ....
______ Quotation marks ....
______ Quotation within quotation........
______ Over punctuation X

C. WORDS AND SENTENCES:
______ Singulars and plurals...... 1, 8, 11
______ Case 5, 6, 9, 12
______ Tense 2, 4, 7, 8, 9, 13-15
______ Good usage 3
______ Recognizing sentences...... 16-25

D. GRAMMAR:
______ Vocabulary 1-7
______ Parts of sentences............ 8-10
______ Kind of sentences............ 11-13
______ Parts of speech............ 14-30
______ Nouns ....
______ Pronouns ....
______ Verbs ....
______ Adjectives ....
______ Adverbs ....
______ Conjunctions ....
______ Prepositions ....

E. SPELLING:
______ ...... very poor ......
______ ....
______ ....
______ ....

F. HANDWRITING:
poor Quality and legibility poor

(This and the material reproduced on the preceding page, copyright, 1934, by E. W. Tiegs and W. W. Clark. By permission of the Southern California School Book Depository, Los Angeles, California, and the authors.)

Table 36 A total scores for *Capitalization, Punctuation, Words and Sentences, Grammar, Spelling,* and *Handwriting* and a total language score are given in profile form for this same test. It appears that James Brown was considerably above average in *Punctuation* and *Grammar,* near the average in *Capitalization* and *Words and Sentences,* and far below average in *Spelling* and *Handwriting.*

b. Measures of Spelling. Investigations of spelling difficulties, which are fairly numerous, are typified by the reports of Gates (1922) and Wheat (1932). Both the studies agree that carelessness is a major source of error and that great improvements come from well-motivated drills. The most common types of errors are:

1. Phonetic – *Wensday* for *Wednesday, hite* for *height*
2. Use of the Vowels – *acheive* for *achieve*
3. Double Letters – *leter* for *letter*
4. Omissions of Silent Letters and Central Syllables
5. Substitution – *goiny* for *going*
6. Mispronunciations – *chimley* for *chimney*

All except the last of these types of errors are probably due to faulty visual memory. From the work of Thurstone (1938), spelling seems to be a special ability. He found an independent mental factor, which he called *W,* and tentatively named Word Fluency in dealing with printed words (see Chapter 11).

For testing purposes the standard scales are nearly all of the survey variety in which items of increasing difficulty are presented. A typical example is the Ayres (1915) scale, which was made from a list of 1,000 common words. After the words had been presented for spelling to large grade groups, the results were tabulated to show the per cent of each grade which succeeded in spelling each word. Words of approximately equal difficulty were grouped together as shown in Table 37, where levels G, O, and Y are presented along with the per cents of words spelled correctly by each grade.

TABLE 37

Buckingham Extension of Ayres Spelling Scale

	Levels		
Grade	*G*	*O*	*V*
II.	84	27	0
III.	94	50	2
IV.	99	73	8
V.		84	16
VI.		92	27
VII.		96	42
VIII.		99	58
IX.			73

G	*O*			*V*
by	eight	remain	lemon	decision
have	afraid	direct	laughter	principle
are	uncle	appear	lying	accommodate
had	rather	liberty	mountains	accuracy
over	comfort	enough	nails	counterfeit
must	elect	fact	needle	dessert
make	aboard	board	nobody	digestible
school	jail	September	oar	immense
street	shed	station	palace	leopard
say	retire	attend	penny	marmalade
come	refuse	between	pitcher	millionaire
hand	district	public	regular	mucilage
ring	restrain	friend	repeats	orchestra
live	royal	during	reprove	parliament
kill	objection	through	sailor	perceived
late	pleasure	police	sentence	possess
let	navy	until	shining	precipice
big	fourth	madam	surface	recommended
mother	population	truly	sweeping	resemblance
three	proper	whole	sweeps	restaurant
land	judge	address	thief	seized
cold	weather	request	waist	superintendent
hot	worth	raise	waiting	surgeon
hat	contain	August	weary	thoroughly
child	figure	Tuesday	writing	
ice	sudden	struck		
play	forty	getting		
sea	instead	don't		
bread	throw	Thursday		
come	personal	canoe		
eats	everything	captain		
food	rate	cellar		
	chief	clothes		
	perfect	covered		
	second	creature		
	slide	curtain		
	farther	declared		
	duty	distance		
	intend	double		
	company	explain		
	quite	fields		
	none	floated		
	knew	holiday		

(Buckingham, 1927. By permission of the Public School Publishing Co.)

Methods of examining one's spelling ability include both recall and recognition. Dictation is one of the commonest methods and one of the most searching, since one has greater difficulty in recalling than in recognizing facts. The Stanford Achievement Test attempts to measure usual spelling habits by not telling the students that the dictation passage is a spelling test.

Another method of examination presents a printed series of words and asks the student to correct those that are misspelled. The Iowa Placement Examinations, Series E. T. I. (1925), present 50 words, of which 25 misspelled words are to be correctly written. The first ten of these are:

________________ acceptance
________________ appreciate
________________ begining
________________ confirming
________________ crocheting
________________ disagreeable
________________ experiance
________________ evidantly
________________ niece
________________ genuine

A third method of examination is illustrated by the Columbia Research Bureau English Test (1926). Here a word is printed four times, once correctly spelled and three times incorrectly, thus:

(*a*) 1. fifty, 2. fivety, 3. fifety, 4. fivty
(*b*) 1. wissdom, 2. wisedom, 3. wisdom, 4. wisdome
(*c*) 1. vanety, 2. vanity, 3. vinety, 4. vanaty

The task is to select the correctly spelled form.

A fourth method of examination attempts to parallel dictation by presenting words written in a phonetic style, and then asking the examinee to give the approved English spelling; thus, *espeshally* for especially and *biznes* for business.

Research is needed to evaluate the relative advantages of these methods of examination. If enough items of a carefully graded sort are given, the results of one type of examination usually correlate very highly with the others.

3. MEASURES OF READING ABILITY. Since the variety of reading activities is large, many types of reading tests have

been constructed. Nearly all reading tests, however, involve four language factors which seem to be somewhat independent. Two of these have just been discussed — namely, knowledge of vocabulary and of techniques of usage and spelling. The other two factors are often called perceptual span and inference. The most common tests of reading, called general tests, demand all of these four factors in unknown amounts; but it is possible, in diagnostic testing, to control three factors and then to measure the fourth in fairly pure form. The analytical approach is important when a person is deficient in only one or two aspects, since it is desirable to provide a special remedy. Both general and diagnostic tests will be described.

Reading test situations are complicated by the fact that both speed and accuracy, which are often incompatible, are frequently desirable aspects of achievement. Since it is very difficult to hold either accuracy or speed constant, the evaluation of these two aspects is a persistent problem.

a. General Reading Tests. Silent reading tests have usually taken the form of a series of sentences or paragraphs, each of which is followed by one or more questions. As the test progresses the items become longer and more involved, the vocabulary harder, and the questions more complex. Table 38 shows an illustration from the Van Wagenen (1933) Unit Scales of Attainment in Reading. The score for this test, which consists of forty items, is the number of items correctly marked in forty-five minutes. No attempt is made to secure separate scores for various types of questions, but questions of both fact and inference are asked. Tests very similar to this one are found in practically all the general achievement tests now published.

b. Diagnostic Tests. Diagnostic tests will be considered under three headings: (*a*) tests of simple comprehension, (*b*) tests which emphasize reasoning or organization of ideas, and (*c*) tests which emphasize perceptual speed and span. The first two of these are listed in Table 36 B with subhead-

TABLE 38

Sample from the Unit Scales of Attainment

Reading

Directions: Read sample paragraph A carefully.

A	A	
America was discovered for Spain by an Italian sailor, Columbus, in 1492. Shortly after this another Italian sailor, John Cabot, sailing from England, reached the coast of Labrador. Still later the country was explored by the French sailor, Cartier, and the Spanish explorers, Cortez and Ponce de Leon. America could be reached only by a long and dangerous voyage across the Atlantic ocean in small wooden sailing vessels.	*1.* The paragraph is mainly about **1. Columbus 2. Labrador** **3. crossing the Atlantic** **4. discovery and exploration of America 5. Spain**	*1.* ..**4**..
	2. By whom was America discovered for Spain? **1. John Cabot 2. Cartier** **3. Columbus 4. Cortez** **5. Ponce de Leon**	*2.* .....
	3. The early explorers came to America in **1. steamboats 2. sail-boats** **3. airplanes 4. canoes** **5. rowboats**	*3.* .....
	4. The early explorers of America were **1. foolish 2. weak 3. cowardly 4. brave 5. fearful**	*4.* .....

Beginning with paragraph **1** below, read each paragraph carefully, then draw a line under the **one** word or phrase that is true for each question or statement at the right of the paragraph and put its number in the space at the right as in the samples above. You may read the paragraph **more than once** if you need to do so.

(Van Wagenen, 1933. By permission of the Educational Test Bureau, Minneapolis, Minn.)

ings which yield a detailed analysis of James Brown's record. It appears from the underlined numbers, which indicate errors, that he did fairly well on simple comprehension items but very poorly on organization and interpretation items. Illustrative items will be given from various tests.

(1) Simple Comprehension. Here the task is to answer questions about facts stated in the context. Thus, the Sangren-Woody (1927) Reading Test, Part III, consists of *Fact Materials* and Part VI, of *Following Directions.* Portions of these are printed in Table 39.

TABLE 39

Samples from the Sangren-Woody Reading Test

Part III. Fact Material

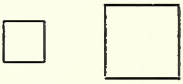
Directions: Write the answer to each question on the dotted line. Use one word if possible.

The "lead" in your pencil is not made of lead. Long ago people had lead in their pencils; that is probably why the pencils we use are called lead pencils. Another mineral called "graphite" is now used. This mineral is taken from mines in the same way as coal or iron ore.	1. What did people use in their pencils long ago? . . . . 2. What mineral is used in the pencils now? . . . . 3. From what is the mineral taken? . . . .

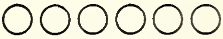
Part VI. Following Directions

Directions: Do what each paragraph tells you to do.

1. At the right are two squares of different sizes. The larger square is a playground for children, and the smaller one is a garden into which children must not go. There should be a fence between the playground and the garden. Make this fence by drawing a line to separate the squares.



2. At the right are six circles. They stand for one-half dozen eggs. The second egg in the row is not a good one and cannot be used for cooking. In order that Mother will not make a mistake and use it, you must take your pencil and mark it with a cross.



(Sangren and Woody, 1927. By permission of the World Book Co.)

(2) Verbal Reasoning and Organization. Some of the tests in this group ask one to draw inferences from a paragraph. These are illustrated by the Gates (1926) Silent Reading Tests which yield scores for: (*a*) appreciating general significance, (*b*) predicting the outcome of given events, (*c*) understanding precise directions, and (*d*) noting details. The last two of these seem to emphasize comprehension of stated facts, but the first two, see Table 40, require inferences as well. These cannot be considered to be pure reasoning tests because the vocabulary becomes more difficult as the test progresses.

TABLE 40

Samples from the Gates Silent Reading Test

TYPE A. TO APPRECIATE THE GENERAL SIGNIFICANCE

This is to be a reading test. You are to read a number of paragraphs. Below each paragraph are five words. One of the words tells how some one described in the paragraph felt — whether sad or happy, etc. You should draw a line under that one — and only one — word to show that you understand just how the person described in the paragraph did feel. Now let us try a sample before we begin the real test. Read the following paragraph and then draw a line under the word which you think tells best how the person felt.

Once upon a time a young fairy went down to the river to swim. She jumped in with a splash. She put out her hands and tried hard to swim. Something seemed to be dragging her down. Oh, it was her wings! She had forgotten to take them off. Fairy wings become heavy when they are wet. She cried for help as loudly as she could.

Draw a line under the word which tells how the fairy felt.

cross angry weary afraid joyful

TYPE B. TO PREDICT THE OUTCOME OF GIVEN EVENTS

This is to be a reading test. You are to read a number of paragraphs. Below each paragraph are four sentences. Each sentence tells what is most likely to follow after the happenings that are described in the paragraph. You should draw a line under one — and only one — of these sentences to show that you can tell what will probably happen next. Now, let us try a sample before we begin the real test. Read this paragraph and then draw a line under the one sentence which you think tells what will happen next.

The grocery man had a black cat. He loved his cat very much. One day a lady brought a big bulldog into the store. The grocer's cat raised his back and said "Meow! Psst!" to the bulldog. Of course, the dog did not like that, so he growled loudly. Before the grocery man or the lady knew what was happening, the bulldog had sprung upon the cat.

They let the fight go on
The cat slept on
The lady took her bird away
The grocery man saved his cat

(Gates, 1926. By permission of the Bureau of Publications, Teachers College, Columbia University.)

The purest form of a verbal reasoning test is thought to be syllogistic (see Table 14). Although such forms are very rarely found in achievement test batteries, they have the advantage of being relatively free from vocabulary variations.

A very hard syllogistic test can be made in which only a Grade III vocabulary is used.

Tests of reorganization use material which has been disarranged with instructions that it be properly arranged. Disarranged sentences are shown in Table 41 A and disarranged paragraphs in Table 41 B. Analysis of factors leading to success in these tests is very difficult when the vocabulary load becomes greater as the reorganization becomes harder. One person may fail the test because of a poor vocabulary; another, because of poor inferences.

Tests of organization of material are illustrated in Table 41 C where a paragraph is printed with its phrases numbered consecutively. The student is asked to select the most important items in the paragraph and arrange them in an outline. The same type of test without numbered phrases is much more difficult to score and probably requires more organizing ability than the numbered form. Much research is needed to determine and appraise varieties of organizing ability.

(3) Perceptual Speed and Span. The amount of written material which a person comprehends at a glance is considered to be an important aspect of reading. Rough techniques for measuring perceptual speed are illustrated by flash cards which are used in the earlier grades. At later ages two standard tests are common: (*a*) word and number comparison tests, and (*b*) speed of reading tests.

Word and number comparison tests are well illustrated by the Minnesota Test for Clerical Workers (1933), which is shown in part on page 23. The score, which is the number items compared, depends upon perceptual speed and familiarity with words and numbers.

Speed-of-reading tests ask one either to detect errors in a passage, or simply to read for a given time and to note the number of words which seem to have been comprehended. The first is illustrated by the Chapman-Cook Speed-of-Reading, which consists of easy paragraphs of 30 words each. Near the end of each paragraph, one word spoils the meaning by its incongruity. The task is to cross out this word. For example, Chapman (1924):

> It was such a cold, boisterous, and wintry day that every person who was walking wore the thinnest clothes that he could find in his clothes closet.

This type of test has been produced by several authors, and Eurich (1931) followed the general idea when he composed a speed-of-reading test at the college level using longer paragraphs and harder words. Criticism has been directed against this type of test since it does not resemble usual reading activities in the following respects:

1. It consists of disconnected passages which do not allow usual rhythms.
2. It requires specific search for a single word rather than a comprehension of phrases.
3. It requires the crossing out of a word.

The other type of speed-of-reading test presents a passage for continuous normal reading during a few minutes but usually fails to control comprehension in such a fashion that

TABLE 41

Samples from the Iowa Silent Reading Test

Advanced Examination, Grades 7–12

A.

TEST 4. SENTENCE ORGANIZATION

(Time Allowance: 4 minutes)

Directions to the Pupil: This test is given to see how well you are able to arrange groups of words into sentences. Work all the exercises as shown in the sample.

Samples: (1) a wagon, (2) a boy, (3) had .....*2, 3, 1*.....

(1) how small, (2) see, (3) he is .....*2, 1, 3*.....

---

1 (1) wanted, (2) to go home, (3) the boy ....

2 (1) always, (2) be rewarded, (3) good deeds, (4) should ....

3 (1) as children, (2) they, (3) get stronger, (4) grow older ....

B.

TEST 5. PARAGRAPH ORGANIZATION

PART C.

(Time allowance: 3 minutes)

Directions to the Pupil: The following exercises are given to test your ability to arrange the sentences of an unorganized paragraph in their proper order. Work all the exercises as shown in the sample.

Sample: (1) One man found that until he put toads in his greenhouse he could not keep insects from eating some of his flowers.
(2) Sometimes men keep toads in their greenhouses. .....*2, 1*.....

The Cattle Tick

1 (1) Once a territory has been made tick free, it is kept so by a strict quarantine against the introduction of infested animals. (2) The cotton states and the Federal government have made great progress in fighting this pest through the system of dipping cattle to rid them of the ticks. (3) Chief among the parasites of cattle is the Texas fever tick, which has caused enormous losses in the South. ......

Early French Explorers in America

2 (1) From this highly strategic post scores of explorers departed to become the pioneers of France in the new world before Boston and Philadelphia had been founded in the English settlements. (2) Quebec, founded by Champlain a year after Jamestown, is located on the St. Lawrence, eight hundred miles from the edge of the continent. (3) The early history of the Great Lakes region is the record of these French explorers. ......

C.

TEST 5. PARAGRAPH ORGANIZATION

PART B. OUTLINING

(Time Allowance: 3 minutes)

Directions to the Pupil: The following are exercises to test your ability to organize an outline giving the most important items of a paragraph. Read the following paragraphs carefully. At the right of the paragraphs are outlines partially filled in. Fill in the blank spaces in the outline from your reading of the paragraphs by placing in the outline the numbers corresponding to the brackets and be sure to select the group of words in the different brackets which will result in a well organized outline. Be sure not to include more items in the outline than have been provided for.

---

1. $\underbrace{\text{In the United States,}}_{1}$ which is $\underbrace{\text{the leading agricultural country}}_{2}$ $\underbrace{\text{in the world,}}_{3}$ $\underbrace{\text{several causes have combined}}_{4}$ to encourage $\underbrace{\text{this industry.}}_{5}$ $\underbrace{\text{Of these factors,}}_{6}$ $\underbrace{\text{the more important}}_{7}$ are the $\underbrace{\text{fertility of the soil,}}_{8}$ $\underbrace{\text{the variety of climate and other conditions of environment,}}_{9}$ $\underbrace{\text{the energy of the people,}}_{10}$ the $\underbrace{\text{encouragement lent by the government to scientific agriculture,}}_{11}$ and $\underbrace{\text{the unrivaled transportation system for marketing}}_{12}$ $\underbrace{\text{crops.}}_{13}$ $\underbrace{\text{Land has been very cheap.}}_{14}$ $\underbrace{\text{High wages in other industries}}_{15}$ $\underbrace{\text{have}}_{16}$ $\underbrace{\text{led to the invention of machinery}}_{17}$ by which $\underbrace{\text{one man can do the work of many.}}_{18}$ There is $\underbrace{\text{no country in the world}}_{19}$ where $\underbrace{\text{machinery is used so extensively in agriculture as in the United States.}}_{20}$

Paragraph 1.

I. Factors encouraging American agriculture.

A. .....

B. .....

C. .....

D. .....

E. .....

F. .....

G. .....

(Jorgensen and Greene, 1927. By permission of the Bureau of Educational Research and Service, University of Iowa and the World Book Co.)

the scores are comparable with one another. Although directions commonly say to read slowly enough to understand what one reads, still some students interpret this to mean skimming and some, a detailed analysis. Tests of comprehension usually follow the reading of a passage, but these introduce factors of recall and possible distraction, and no very satisfactory way of combining speed and comprehension scores has appeared.

Two very elaborate and expensive techniques for measuring perceptual span of words should be mentioned. One technique, which uses a Metronoscope, presents words or phrases to a student at a given rate and asks him to record or tell what he has seen. The other technique photographs eye movement while one is reading. The first method does not resemble normal reading in some respects, but it does give accurate records under the test conditions. It has been proposed as a means of training slow readers. The second method usually fails to evaluate comprehension, but it does show precise records of number, duration, and order of fixations on a line.

A less expensive technique for training readers has been described by Dearborn and Anderson (1938). Printed material was photographed on motion picture film in such a way that when the film is projected, successive phrases and lines are exposed as they have been grouped by skilled readers. The exposure times are limited to approximately one-fifth of a second, thus preventing more than an optimum number of fixations per phrase. The suddenness with which the units appear and disappear controls the duration of fixation time. At the beginning of training the material is simple, the phrases short, and the exposure rate slow. As training progresses, the material increases in complexity and the projection rate is more rapid. Training of this sort has resulted in considerable improvement as shown by the Minnesota Speed of Reading Tests and the Gates Silent Reading Tests on small samples of both college students and elementary-school children. Although standard tests of reading ability using films of this sort have not yet come to hand, the technique may be readily adapted to examinations. Research is needed to devise measures of speed-of-reading which will be typical of actual reading situations and free from ambiguous results.

III. MEASUREMENT OF MATHEMATICAL ABILITY

Tests in this group range from the simplest counting operations to highly complex problem solving. There are three types of skills involved which appear to be highly independent, at least among adults:

1. Calculation from rote memory and rule
2. Abstract reasoning with numbers or letters
3. Geometric reasoning with spatial data

A. Arithmetic Tests

A very popular type of general arithmetic test, shown on page 300, is called arithmetical reasoning or problem solving. Tests of this sort require the three factors mentioned above in unknown combinations and some language ability as well. Nearly all of the elementary tests contain many interesting practical applications from business, architecture, surveying, and house management. Tests of this general sort are of little value for a diagnosis of basic skills, but they do show the results of complex thinking about common objects.

A diagnostic test of mathematical ability is reported in Table 36 A. Here eight subdivisions are listed, and it appears that James Brown is above average in *Problems, Number Concepts, Symbols and Rules,* and *Numbers and Equations,* and near or slightly below average in *Addition, Subtraction, Multiplication,* and *Division.* Table 36 B shows, with its detailed analysis, that his most difficult items in all eight divisions demanded skill in using fractions and decimals. He was able to handle simple and abstract numbers fairly well.

The Progressive Achievement Test and a number of similar tests are particularly useful for a rapid survey. The subdivisions are probably too short for very reliable testing such as is needed for appraising experimental education. Very elaborate instruments are to be found in the Compass Diagnostic Tests, Ruch *et al.* (1925). Approximately 90 distinct acts were listed, and a large number of separate tests were printed. One of these, Addition of Fractions, is shown in Table 42. Furthermore, a Standard Arithmetic Work Book was designed to furnish both original learning situations and remedial drills. Such work books are now numerous and widely used in both traditional and progressive schools.

TABLE 42

Samples from the Compass Diagnostic Tests in Arithmetic

Test V: Addition of Fractions and Mixed Numbers — Form A

PART 1 — CHANGING FRACTIONS TO EQUIVALENT FORMS

Directions: Change the form of each fraction below to use the denominator given. Study the samples carefully to see how this is to be done.

Samples: $\frac{1}{8} = \frac{2}{16}$ $\frac{2}{3} = \frac{8}{12}$ $\frac{1}{3} = \frac{}{9}$

$\frac{1}{2} = \frac{}{12}$ $\frac{1}{2} = \frac{}{10}$ $\frac{2}{5} = \frac{}{10}$ $\frac{1}{2} = \frac{}{4}$ $\frac{4}{9} = \frac{}{36}$ $\frac{1}{12} = \frac{}{36}$ $\frac{4}{9} = \frac{}{18}$

$\frac{5}{6} = \frac{}{18}$ $\frac{1}{2} = \frac{}{6}$ $\frac{2}{3} = \frac{}{6}$ $\frac{3}{4} = \frac{}{44}$ $\frac{3}{5} = \frac{}{20}$ $\frac{1}{2} = \frac{}{20}$ $\frac{1}{8} = \frac{}{24}$

PART 4 — FUNDAMENTALS OF ADDITION OF FRACTIONS

Directions: Find the sum for each example below. Do your figuring on a piece of scratch paper.

$\frac{5}{8} + \frac{3}{8} =$ $\frac{8}{20} + \frac{3}{5} =$ $9 + \frac{2}{5} + 11\frac{3}{35} + 16\frac{6}{7} =$

$5\frac{7}{12} + \frac{1}{2} =$ $\frac{1}{8} + 8\frac{5}{6} + 2\frac{1}{3} =$

$$\begin{array}{r} 4\frac{2}{5} \\ 9 \\ \frac{1}{2} \\ 16 \\ \frac{2}{4} \\ \hline \end{array}$$

Three-tenths plus one-half = Add $\frac{1}{7}$ and $\frac{1}{11}$ =

(Ruch, Knight, Greene, and Studebaker, 1925. By permission of Scott, Foresman and Co.)

Another type of analysis is illustrated by the use of the Buswell-John (1925) Diagnostic Chart. A trained observer watches a pupil work through a standard eight-page test and notes the types of errors as they are made. Table 43 shows a check list of errors in addition. Similar lists are available for multiplication, division, and subtraction. Direct observation gives more insight into the sources of error than unobserved test scores, particularly if the student talks as he works.

B. Algebra, Geometry, and Trigonometry Tests

For both high school and college, general tests of algebra, geometry, and trigonometry have been well standardized, and methodical work books are available for the elementary courses. Thorndike *et al.* (1923) wrote an elaborate *Psychology of Algebra* which listed fourteen algebraic abilities, suggested ways of eliminating unnecessary habits, and also provided or suggested drill on neglected skills.

Out of a large number of available algebra tests two will be mentioned. Lee (1930) designed a test to indicate ability to succeed in algebra. Its four parts are:

a. Arithmetic problems where algebra might be helpful
b. Number analogies, such as: 3 – 9::5 – 30, 15, 20, 23
c. Number series: 4, 8, 10, 20, 22, 44_____
d. Easy formulas: $A = \frac{ab}{2}$. Given a = 10, b = 8, to find A

Lee reported correlations of .71 among 318 high-school pupils between this Algebraic Ability Test and an Achievement Test after one semester of algebra. This correlation is nearly the same as that between Achievement Tests at 9 and 18 weeks. The Ability Test correlated .631 with algebra grades. These figures are considered to be about as high as the reliability of grades will permit.

The commonest type of Algebra Achievement Test is illustrated by the Columbia Research Bureau Algebra Test,

TABLE 43

Teacher's Diagnostic Chart for Individual Difficulties

Fundamental Processes in Arithmetic

Teacher's Diagnosis
for pupil _____

Name____________________ School____________ Grade___ Age___ IQ___

Date of Diagnosis:______________ Add.___; Subt.___; Mult.___; Div.___

Teacher's preliminary diagnosis ______________________________________

ADDITION: (Place a check before each habit observed in the pupil's work)

_____ a1 Errors in combinations
_____ a2 Counting
_____ a3 Added carried number last
_____ a4 Forgot to add carried number
_____ a5 Repeated work after partly done
_____ a6 Added carried number irregularly
_____ a7 Wrote number to be carried
_____ a8 Irregular procedure in column
_____ a9 Carried wrong number
_____ a10 Grouped two or more numbers
_____ a11 Splits numbers into parts
_____ a12 Used wrong fundamental operation
_____ a13 Lost place in column
_____ a14 Depended on visualization
_____ a15 Disregarded column position
_____ a16 Omitted one or more digits
_____ a17 Errors in reading numbers
_____ a18 Dropped back one or more tens
_____ a19 Derived unknown combination from familiar one
_____ a20 Disregarded one column
_____ a21 Error in writing answer
_____ a22 Skipped one or more decades
_____ a23 Carrying when there was nothing to carry
_____ a24 Used scratch paper
_____ a25 Added in pairs, giving last sum as answer
_____ a26 Added same digit in two columns
_____ a27 Wrote carried number in answer
_____ a28 Added same number twice

Habits not listed above ____

____

(G. T. Buswell and Lenore John, 1925. By permission of the Public School Publishing Co., Bloomington, Ill.)

which has two parts. In Part I, skill in solving equations is required, such as:

1. $X + 15 = 23$
14. $4X^2 - 16X - 9 = 0$
18. sq. rt. of $X - 2 = 6$

In Part II, problems are to be solved, such as:

2. How long must I make a garden that is 8 feet wide so that it will have as large an area as my neighbor's, which is 16 feet long and 6 feet wide?
24. The distance (S) a body falls in T seconds is expressed by the formula $S = 16T^2 + VT$, in which V is the initial velocity downward. How many seconds will it take a body to fall 640 feet if thrown downward with a velocity of 48 feet a second? (Use T for the unknown.)

The scores in Part II involve specific information outside of the field of algebra, and some of the problems can be solved without algebra. A great deal of work is now under way to prepare algebra tests which will place greater emphasis on utility and cultural aspects by the use of more challenging problems and the simplest techniques. Yielding to continuous agitation for curricula based on immediate needs and interests of students, many high schools have in recent years transformed the teaching of algebra from the techniques of solving polynomials to approaches to analytical thinking about practical situations, as tested in Part II.

A good illustration of an elementary Geometry Achievement Test is that of Schorling and Sanford (1926), part of which is shown in Table 44.

Part I. A vocabulary test
Part II. Requires a statement of what can be proved from given data
Part III. Asks that the correctness of certain conclusions be judged
Part IV. Requires an analysis of constructions
Part V. Demands the computation of distances or areas

TABLE 44

Samples from the Schorling-Sanford Achievement Test in Plane Geometry

Form B

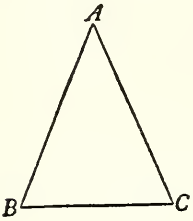
Directions for Part II

Drawing Conclusions from Given Data

In Part II of this test, you will be given geometric figures and certain facts relating to each. Think what fact you could prove about each figure on the basis of the given information and write a statement of this in the space beside the words "We can prove that."

Be sure to use each piece of information that is given to you. You will be allowed 12 minutes for Part II.

Example:



Given ΔABC with $AB = AC$.

We can prove that......................

[Since ABC is given as an isosceles triangle, we can prove that $\angle B = \angle C$, so

$$\angle B = \angle C$$

should be written on the dotted line.]

Directions for Part V

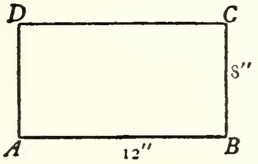
Computation

In Part V of this test, you will be given diagrams about which certain facts are known. You are to use these facts in computing such things as the length of a line, the size of an angle, etc., according to the questions that are asked. Read each question, study the figure upon which the question is based, and write the answer in the space beside the question.

You will be allowed 12 minutes for Part V.

Example:

In the rectangle $ABCD$, AB is 12 in., BC is 8 in.



What is the area of the rectangle? Ans. ....
[*96 sq. in.* should be written on the dotted line.]

Now try this:

What is its perimeter? Ans. ....

Be sure to express your results in *numerical form.*

(Schorling and Sanford, 1926. By permission of the Bureau of Publications, Teachers College, Columbia University.)

Total scores on tests of this sort usually correlate approximately .75 with the midyear grade from a beginning course. They also show equivalent-form reliability of about .85. The subdivisions, which are doubtless less reliable, furnish valuable evidence of a pupil's weak and strong points. The intercorrelations of the subdivisions are low enough, usually about .50, to suggest a number of independent skills, but no careful factor analysis of such skills has come to hand. A subjective analysis indicates that probably four independent factors are at work: verbal information, reasoning, spatial imagery, and number relations.

Tests of advanced work in algebra and in plane and solid geometry are found in the Cooperative Test Series. These include items to appraise calculations and reasoning as emphasized in the standard textbooks.

A good illustration of a standard trigonometry test is that of the American Council on Education Trigonometry Test (1930). There are five parts which are to be completed in 72 minutes.

I. Technical knowledge, as:
   Sec O is positive and tan O is negative, the quadrant of O is: first, second, third, fourth
II. Completion of equations:
   X = 1, Y = 3; tan O = __________
III. Indicate method of solving verbal problem:
   From the top of a cliff 350 feet high the angle of depression of a buoy is 16° 38′. Find the distance from the buoy to the top of the cliff
IV. Indicate method of solving diagrammed problems similar to Part III
V. Like Part IV, but more difficult

IV. SPECIAL STUDIES

Achievements in special studies have been appraised with local tests more often than with standardized instruments. This is probably due to the variation in objectives in different localities and the lack of well-standardized, acceptable tests. Most of the published tests cover only elementary aspects since advanced activities are extremely complex and are carried on by only a few persons. Special tests of Physical Sciences, Social Studies, Foreign Languages, and Business will be briefly discussed.

A. Physical Sciences

Appraisals of the results of any science course have traditionally taken the form of questions about facts and natural laws. A good illustration is the Iowa High School Content Examination, part of which appears in Table 45. It consists of 50 multiple-choice questions on a variety of subjects. A more diagnostic procedure is shown by tests of the Cooperative Test Service. This nonprofit organization, established by the American Council on Education in 1930, has produced, through a large number of contributors, annual forms of reliable tests of English, foreign languages, mathematics, natural sciences, and social studies for high-school and college levels. The Cooperative Chemistry Test (1934), illustrated in Table 4, page 26, is interesting because it attempts an analytical approach by the use of three parts. Part 1 requires knowledge of facts and natural laws, Part 2 requires knowledge of terminology, and Part 3 requires statements of probable outcomes and also explanations for these statements. Part 3 is probably much more dependent upon reasoning activities than Parts 1 and 2. Tests of the type of Part 3 are being more extensively used now than formerly because they are thought to appraise an important objective of the course — namely, insight into causal relationships. Buckingham and Lee (1936) found marked differences between tests of science facts and tests of ability to point out relations between laws and phenomena. Some of their freshman group made good scores on the factual tests and very low scores on the tests which demanded careful inferences. Their results, which are typical of several other

TABLE 45

Samples from the Iowa High School Content Examination

Section 3

SCIENCE

(Time allowed: 10 minutes)

1. Combustion is another name for (1) freezing (2) drying (3) boiling (4) burning (5) melting .....

2. A gas which supports combustion is (1) hydrogen (2) nitrogen (3) carbon dioxide (4) oxygen (5) carbon monoxide .....

3. $H_2SO_4 + BaCl_2 = 2\ HCl$ and (1) H_2O (2) $Ba(OH)_2$ (3) SO_3 (4) $BaSO_4$ (5) H_2ClO_3 .....

4. The freezing point on the Centigrade thermometer is (1) − 273° (2) 0° (3) 32° (4) 100° (5) 212° .....

5. Substances which hasten a chemical action without themselves undergoing any chemical change are called (1) catalysts (2) electrolytes (3) ionogens (4) allotrops (5) anhydrides .....

6. The formula for hydrochloric acid is (1) $NaOH$ (2) HCl (3) HNO_3 (4) $MgSO_4$ (5) H_2SO_4 .....

7. Hydrogen may be made by the action of hydrochloric acid on (1) zinc (2) sodium chloride (3) copper sulphate (4) sodium hydroxide (5) potassium chlorate .....

8. The length of a meter in inches is about (1) 12 (2) 27 (3) 39 (4) 72 (5) 144 .....

9. An example of a chemical change is the (1) desiccation of spores (2) oxidation of sugar (3) dissolving of salt (4) osmosis of glucose (5) exchange of gases in the lungs .....

10. Barometers are used to measure (1) humidity (2) rainfall (3) air-pressure (4) gravity (5) electricity .....

11. The center of our universe is the (1) earth (2) moon (3) sun (4) Jupiter (5) Mars .....

12. An instrument depending on atmospheric pressure for its operation is the (1) siphon (2) hydraulic press (3) thermometer (4) telephone (5) voltaic cell .....

13. The force needed to raise a weight of 1200 pounds on a hydraulic press whose piston areas stand in the ratio of 1 : 6 is (1) 100 lbs. (2) 200 lbs. (3) 300 lbs. (4) 600 lbs. (5) 2400 lbs. .....

14. One organism which possesses antennae is the (1) earthworm (2) starfish (3) hydra (4) amoeba (5) grasshopper .....

15. Electric charges can be detected by means of (1) condenser (2) electroscope (3) dynamo (4) voltaic cell (5) spectroscope .....

(Ruch, Cleeton, and Stoddard, 1925. By permission of the Bureau of Publications, University of Iowa.)

reports, point to the need of measuring knowledge of facts separately from reasoning when careful analyses are desired. (Similar biology and physics tests at both high-school and college levels are found in the Cooperative Test Series, the American Council on Education Series, and a number of others.) Other objectives of a science course, such as skill in the laboratory, known scientific theories, and practical applications, are very difficult to appraise. Wrightstone (1936) has published an initial study of such appraisals in high-school science courses.

B. Social Sciences

History, civics, and geography are the main social studies in high school, and nearly all the standard tests deal with these topics. Social studies broaden into the humanities in college, including economics, sociology, philosophy, psychology, and political science. A comprehensive volume by Kelley and Krey (1934) discusses tests of knowledge and attitudes in social studies. In this volume the report of Pressey is of particular interest. She first established the master list of terms used in history, shown in Table 32, and then constructed multiple-choice items for 346 words and applied these items to more than 11,000 pupils in Grades IV, VI, VIII, X, and XII. The proportions of students in each grade who passed seven of the items are shown in Figure 34. This figure shows that mastery of three of the items was fairly complete in Grade VIII and that the item for the word *enact* was answered about as well by Grade IV as by Grade X. The items for the words *competition* and *domestic* showed fairly regular increases, and the item for *sedition* proved to be very difficult before Grade X.

The seven items are shown in Table 46. An inspection of these items will show that some words are defined more completely than others. Some items require difficult discriminations in word usage. The difficulty of an item is here determined partly by difficulty of defining the key

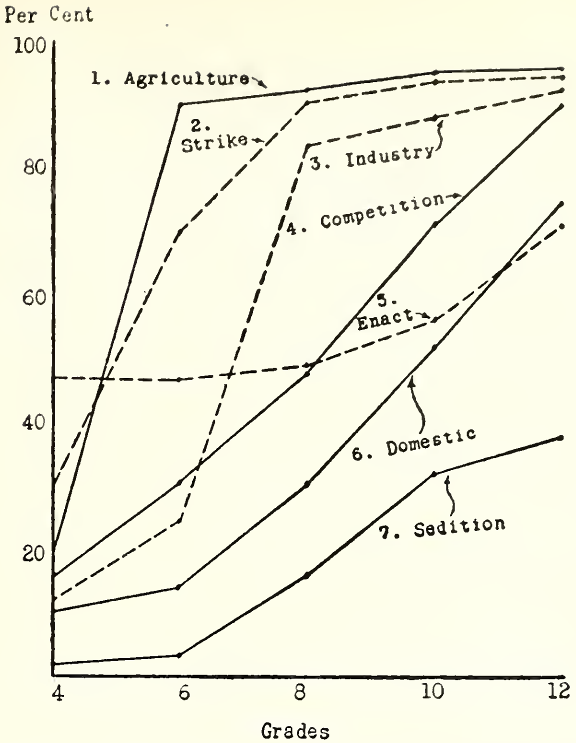


FIGURE 34. **Growth of Social Vocabulary**

Per cents of students in grades IV to XII who succeeded on certain vocabulary items.

(After Pressey, 1934. By permission of Charles Scribner's Sons.)

TABLE 46

Test of Concepts Used in the Social Studies

1. To what does agriculture refer? (*a*) fishing (*b*) mining (*c*) farming (*d*) manufacturing
2. How do laborers most often attempt to force an employer to raise their wages? (*a*) by going to war (*b*) by going to another factory (*c*) by going to a foreign country (*d*) by going on strike
3. Which phrase refers to the entire business of making woolen clothing? (*a*) the woolen shops (*b*) the woolen industry (*c*) the wool-growers (*d*) the sheep ranchers
4. Which of the following words applies to the practice of lowering of prices by one company so as to get business away from another? (*a*) rebellion (*b*) emigration (*c*) output (*d*) competition
5. Which are enacted? (*a*) verdicts (*b*) debts (*c*) laws (*d*) dispatches
6. Which word refers to the affairs relating to one's own country? (*a*) foreign (*b*) international (*c*) domestic (*d*) diplomatic
7. Which is a form of conspiracy against one's country? (*a*) sedition (*b*) opposition (*c*) criticism (*d*) immigration

(Pressey, 1934, p. 189. By permission of Charles Scribner's Sons.)

word, and partly by the context of the item. In order to discover whether such multiple-choice items are indicative of one's ability to write a correct definition, Pressey made a comparison of scores on two vocabulary tests, using the same words. In one test the children checked multiple-choice items, such as those in Table 46, and in the other, they wrote definitions for the key words. She found that approximately 70 per cent of the words correctly defined were also used correctly in the multiple-choice forms and that the correlation between total scores on the two tests was very high. The form in which an item is cast undoubtedly affected the task to some degree, but the results from the two types of tests were very similar. The relative advantages and the intercorrelations of various forms of test items have been discussed in Chapter 6.

From studies such as this the effectiveness of various learning situations may be compared for each item. Pressey concluded from her study (p. 174) that "not enough special vocabulary is acquired for the reading of the average textbooks used in these grades." From such studies it is also possible to compose tests of items which have known levels of difficulty for particular populations. Kelley and Krey (1934) suggested that the best items for careful scale construction were those which, like Nos. 4 and 7 in Table 46 and Figure 34, show regular growth increments. Standard high-school tests of this sort are the Wesley (1932) Tests of Social and Political Terms, the Kelty-Moore (1934) Tests of Concepts in Social Studies, the Brown-Woody (1928) Civics Tests, and several others. College tests are represented by the American Council on Education Civics and Government Test (1929) and the annual Cooperative Tests of American, Ancient, Modern European, English, and Medieval History, of Economics, and of Contemporary Affairs. Terminology, descriptive information, and careful inferences are measured separately, and also added to give a general score.

A number of geography tests have been prepared by state testing bureaus. Typical of a general test is that of Torgensen (1933). This is a sixty-four-item multiple-choice test for Grades VI to VIII which requires knowledge of the earth surface, tides, natural resources, populations, climates, and map reading.

A test of general skills that are basic to effective work in social sciences was published by Wrightstone (1936) in four parts, each of about 25 items:

Part 1. Obtaining Facts: measures a pupil's ability to read charts, graphs, and maps and to use an index
Part 2. Organizing Facts: a test of outlining materials and judging their reliancy
Part 3. Interpreting Facts: a reading test in which statements are to be marked true or false according to a given paragraph
Part 4. Applying Generalizations: using multiple choices, as in Table 38, to explain events and to support the explanation

A test very similar in design and content was also published for natural sciences by the same author. Tests of this sort have not yet been widely enough used to be evaluated. They undoubtedly tap the same sorts of skills that have been described under the headings of language and mathematics, but the combination is probably somewhat unique.

An appraisal of Activities in Social Studies, shown by the check list in Table 47, was used by Wrightstone (1936). It divides activities into three groups: self-initiated, co-operative, and recitational. In his comparison of traditional and progressive schools he found that pupils in the latter showed much more self-initiated activity than those in the traditional schools. This check list is a very significant instrument because it draws the attention of instructor and pupil to activities which are probably of much greater social importance than the acquisition of information in a limited field.

TABLE 47

Check List of Activities in Social Studies

A. Self-Initiated Activities
  1. Bringing voluntarily contributions (clippings, exhibits, books, charts, etc.) for school activities
  2. Submitting voluntarily and orally data, or information, gained outside school (observation, trips to buildings, places, travel, museum, radio, lectures, movies, independent reading, etc.)
  3. Presenting an organized written report showing research or investigation by pupil
  4. Volunteering as leader or special worker on project or task
  5. Suggesting methods, materials, activities, etc., for developing a project or problem
  6. Defending a point of view in which the pupil believes

B. Co-operative Activities
  1. Criticizing (praising or challenging) a contribution
  2. Asking chairman or teacher pertinent subject matter questions which relate to the theme or topic of the group discussion, excluding routine class management questions
  3. Offering objects (books, chairs, pencils, etc.) to teacher, pupil, or visitors
  4. Responding quickly to requests for quiet, material, etc.

C. Recitational Activities
  1. Responding to questions on assigned textbook or subject matter

(Wrightstone, 1936, p. 134. By permission of the Bureau of Publications, Teachers College, Columbia University.)

C. Foreign Languages

Carefully standardized tests are available for some of the skills taught in beginning courses in foreign languages. The Cooperative French Test (1934) is a good illustration:

Part I. Consists of 80 multiple-choice items where a short statement in French is to be read and completed sensibly thus:
  1. Le mari de ma grand'mère est mon — grand-père, cousin, père, oncle

Part II. Here are 100 vocabulary items where a key word is to be matched with its closest synonym.
  1. retourner; 1. rentrer, 2. finir, 3. toucher, 4. recevoir, 5. tomber

Part III. Here 100 grammar items appear in which the correct form is to be supplied from 5 choices.

Sample:

1. le — I have a pen. J'ai —— plume. (3)
2. la
3. une — Here is the book. Voici —— livre. (1)
4. un
5. no additional word — She is in France. Elle est en – France. (5)

Tests of this sort have been criticized by French teachers on several counts, such as:

1. The material is too fragmentary.
2. The material is uninteresting.
3. Some vocabulary items can be correctly answered by noting similarities in word forms, without knowledge of French.
4. The student is not asked to compose any French.
5. The test situation is very different from normal reading situations.
6. The grammar section emphasized proofreading ability which shows little relationship to skill in reading or speaking correctly in actual life situations.
7. There is no check on spelling or use of accents.

For survey purposes, however, these tests are usually recognized as valid measures of general vocabulary, proficiency in comprehension of short passages, and the commonest French syntax. The Cooperative Tests Service has developed similar tests for achievement in German and Spanish. Elaborate tests of the same sort are to be found in the College Entrance Board Examinations and in the several state examination programs. Although tests of ability to comprehend spoken words and to speak a foreign language fluently are found locally, they do not appear to have been standardized.

D. Business Achievement

Tests in this field may be conveniently placed into two groups: clerical skills and business information. The former are fairly well developed, whereas the latter are still in an experimental stage because they are more varied and not yet well validated.

1. Clerical Tests. In 1922 Thurstone published an examination in typing in three parts. The first part presents a typewritten page upon which numerous corrections had been made with pen and ink and asks that a correct copy be made. The second part requires 40 hand-written items to be typed in columns on a blank. The third contains 48 words among which misspelling is to be detected. Separate scores for errors and speed are to be recorded and also combined by the addition of total scores. The skillful typists made the lowest scores.

Blackstone (1923) standardized tests of typing in which business letters in correct typewritten form were presented to be copied during a three-minute period. Norms for speed and accuracy of typing were given separately and also combined into a score thus:

$$\text{score} = \frac{\text{strokes-per-minute} \times 10}{\text{errors} + 10}$$

The scores were found to have a repeat reliability of .93 for a group of pupils with 20 months instruction. The average pupil with 5 months instruction secured 88 points, those with 20 months, 206 points, and with 30 months, 236 points.

Shorthand tests have been standardized in connection with courses of instruction. Bisbee (1933) prepared a test which is typical. Test 1 measures skill in outlining approximately 122 different Gregg principles, 60 phrases, and 70 brief forms. In Test 2, two letters are dictated at different speeds. From this material errors in outlining were scored. Test 3 requires correct English spelling of 42 hard words. Transcription of Tests 1 and 2 is also required after their outlines have been scored. The typewritten sheets are scored for errors in transcription, spelling, and punctuation.

A different type of stenography test is that of Blackstone and McLaughlin (1932), which has seven parts. The first six parts are short examinations of English techniques and information about business practices; the seventh, a transcription test. Thus:

1. Grammar, spelling, and punctuation errors were to be detected (30 items, 8 minutes)
2. Correct syllabification was to be recognized from 4 choices (20 items, 4 minutes)
3. Office practices: This is an information test with a few items concerning the right thing to do (20 items, 5 minutes) Thus:
   a. Biographical sketches of men who have been successful in the arts and sciences may be found in:
   1. The World Almanac, 2. The Statesman's Yearbook, 3. R. G. Dun's Reports, 4. Who's Who. (—)
   b. On the first day on a new job, a girl is called by her first name by one of the men working with her. She does not like this. She should:
   1. Tell him that she prefers to be called by her last name, 2. Show her disapproval by a cold manner, 3. Call him by his first name, 4. Report him to her chief. (—)
4. The alphabetical filing of 20 names among 28 already listed alphabetically (6 minutes)
5. Abbreviations of 20 terms are to be written, such as, Certified Public Accountant (C.P.A.), and the unabbreviated expression is to be supplied for 20 abbreviations, such as, cr. (credit), mgr. (   ) (5 minutes)
6. Business organization: In this test, twenty types of information are listed, and twelve department names. The task is to indicate the department which should be consulted for each type of information (5 minutes)
7. Transcription: Here two short letters are read by the examiner and 12 minutes are allowed for typing both letters. The manual provides seven letters arranged in order of difficulty of complexity, and prints the standard speed of reading the short passages in each letter. The more complex letters are to be read faster ranging from 36 to 120 words per minute. The hardest transcriptions are allowed five times as much credit as the easier. The examiner is told to choose the letters which she thinks the class as a whole can transcribe "completely and correctly."

Norms for these seven tests are available for 1000 pupils distributed among four semesters of high-school work, and for a large sample of stenographers. The stenographers and the fourth semester pupils had the same total scores. The latter were 12 points ahead of the former in transcription, but the stenographers showed more skill in tests 2, 3, and 5. The equivalent-form reliability for the total score was reported to be .88 and correlations between scores and efficiency ratings were .62 and .79 for small groups in service.

This test is interesting because it appraises more than transcription ability. Tests 1, 2, and 5 are typical of English usage in general, Tests 3 and 6 deal with office background, and Test 4 requires filing activity, all of which can be done by persons without any stenographic training. Such items have been included in tests of general clerical work. One of the first of these was developed by Thurstone (1922). It consists of eight parts:

1. Checking addition and subtraction of numbers
2. Detecting misspelled words in a long passage
3. Drawing a line through X, Z, U, and C but no other letters in pied type
4. Associating letters with numbers
5. Classifying men's names according to location in cities and also alphabetically
6. Classifying insurance items according to amount and date simultaneously
7. Easy arithmetic problems
8. Matching proverbs

Since many of the skills needed in this test may be well developed before one starts formal clerical training or employment, the Thurstone Clerical Test has often been used as an aptitude test to predict success. A number of clerical aptitude tests have been published along these lines. Information concerning their use and validity should be sought in technical works on Industrial Psychology.

2. Bookkeeping Tests. An elaborate set of four bookkeeping tests at high-school level was prepared by Breiden-

baugh (1940). A Work Sheet, Balance Sheet, Profit and Loss Statement, and Closing Entries are to be filled in correctly from figures that are furnished. Knowledge of terms and business practices is also appraised by separate tests. These tests were constructed to cover the basic practices which are given in several textbooks. Their reliabilities are not given, nor is any indication of prediction of success in bookkeeping work. The tests are designed as achievement examinations in a course of instruction.

A more general test of bookkeeping is that by Elwell-Fowlkes (1928). Without making actual calculations, one is required to give information about general theory, journalizing, classification, adjusting entries, closing the ledger, and statements. Fifty minutes are allowed to complete 90 items. Two forms are available which correlated .82 among 258 students in half-year classes.

3. BUSINESS INFORMATION. A general test of business information was published by Thurstone (1921) and a similar but more detailed test by Thompson (1937). In the latter, 220 items, which are to be completed in 80 minutes, are distributed as follows:

1. Arithmetic	17
2. Communication	26
3. Money and Banking	30
4. Purchasing	11
5. Recording Filing	16
6. Organization and Ownership	30
7. Economics	17
8. Selling and Advertising	19
9. Investment and Insurance	41
10. Travel	13

Norms for 790 high school students are given for the two forms. Total scores on the two forms were found to correlate approximately .94 for three groups of about one hundred students. No figures are given to show the reliabilities of the subtests. The subtests are probably too short to be used for diagnoses of separate skills although they do furnish a rough basis for quantitative analyses.

V. STUDY HABITS

A fairly large number of manuals on how to study effectively are available, and many high schools and colleges have courses and clinics designed to help students diagnose and remedy bad habits. The appraisal of study habits, however, is usually a rough rating either by a teacher or by the student himself. A number of studies of poor and good practices have appeared, of which that of Wrenn (1934, revised 1940) is representative. He secured two groups of students, one from the best ten per cent and the other from the lowest 20 per cent of students in college scholarship, matched on college aptitude test scores, major field, sex, and length of time in college. This left the difference in scholarship as the only major difference between the two groups so that differences found in study habits might be thought of as affecting the difference in scholarship. Both groups answered 69 questions relating to their own study habits and attitudes. The tabulated results showed that 30 of the statements distinguished between the two groups; therefore, these statements were incorporated into Wrenn's *Study Habits Inventory*, part of which is shown in Table 48. It has three groups of items: (*a*) concerning note-taking and reading, (*b*) habits of concentration and interests, and (*c*) general habits of work. The results in (*a*) show that the poorer students are slow readers, have trouble picking out important points, have poor command of fundamental vocabulary and exposition, rely on arbitrary systems, and fail to overlearn material. In (*b*) the poorer students are characterized by poor concentration, marked dislike for certain professors or subjects, inability to work under pressure, day-dreaming, and various dissipations. In (*c*) the poorer students are found to avoid studying alone, to become nervous during examinations, to distribute their time unwisely, to avoid systematic reviews, and to become too tired. Wrenn (1931) suggested various remedial procedures, some involv-

TABLE 48

Study-Habits Inventory

	1 *Rarely or Never*	2 *Sometimes*	3 *Often or Always*	*Score*
A. NOTE-TAKING AND READING TECHNIQUES				
1. I miss important points in the lecture while copying down notes on something which has gone before . . .	......	......	......	......
2. I have trouble picking out the important points in material read or studied: take down material which later on turns out to be merely explanatory or irrelevant	......	......	......	......
3. I have to re-read material several times — the words don't have much meaning the first time I go over them .	......	......	......	......
(*5 more items*)				
B. HABITS OF CONCENTRATION AND SCHOOL INTERESTS				
9. I find it hard to keep my mind on what I am studying — don't know what I have been reading about when I get through	......	......	......	......
10. I have a tendency to "day-dream" when trying to study	......	......	......	......
11. It takes me some time to get settled and "warmed up" to the task of study	......	......	......	......
(*7 more items*)				
C. GENERAL HABITS OF WORK				
19. My study periods are often too short for me to get "warmed up" and concentrated	......	......	......	......
20. I am careful to make proper use of reviews — reviewing notes taken, setting aside time for daily or weekly reviews during quarter, concentrating review mainly on the points where I feel weakest	......	......	......	......
21. I get "fussed" and nervous on exams — blow up and can't do myself justice or tell what I know	......	......	......	......

(*9 more items*)

(Wrenn, 1934. By permission of the Stanford University Press.)

ing improvement in language techniques, and others a thorough reorganization of vocational and social plans with the aid of a mental hygienist.

VI. INDICES OF ACCOMPLISHMENT

Since the raw scores of achievement tests do not show the significance of the finding, several methods of interpreting scores have been devised. One group of investigators have used mean scores of age or grade groups to construct scales. Figure 17, page 63, shows a profile which give both educational age (EA) and chronological age equivalents and mean scores for each tenth of a grade, from grades 2.6 to 10.0.

From this figure it appears that Eleanor Brown's highest score, 100 in *Reading: Word Meaning,* is equivalent to an EA of 15 years, 8 months, and to the mean of Grade 9.7. Her lowest score, 76 in dictation — in this case a spelling test is equivalent to an EA of 12 years, and to a grade standing of 6.2. The educational ages above 15 years and below 7 years were not based upon actual measurement but upon the extrapolation of a growth curve (see Figure 99, page 673) which was found to fit the scores for the intermediate ages.

Educational ages and grade equivalents are valuable for placing or promoting a pupil in schools where promotions are based upon achievement. They do not, however, furnish an indication of where a person is in his own age or grade group, — among those with whom he must compete. This information is furnished by centile scores which are demonstrated in Table 36 A. Here the proportion of pupils that exceed a given pupil can be quickly read. A chart which combines both centile and grade norms for a single test is to be seen in Table 10, page 49. From this table it appears that about 13 per cent of students in the grades below college were retarded one grade; 10 per cent, two grades; and 8 per cent, three grades. Similarly, approximately 14 per cent were advanced one grade; 10 per cent, two grades; and 8 per cent, three grades. Figures of this sort vary according to the method of grouping students which happens to be in effect in a particular school. The spread of age and grade scores in two city schools are contrasted in Figure 35. Although the data refer to intelligence test

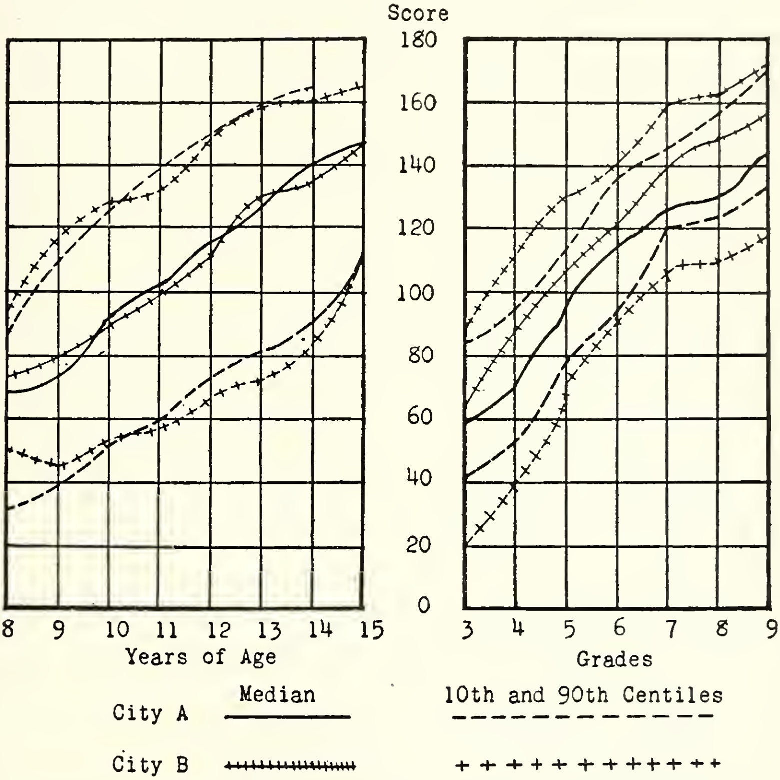


FIGURE 35. **Intelligence of School Children in Two School Systems**

(By permission of Pressey, 1933, and Harper and Bros.)

scores, similar results are found when general educational achievement tests are used. The medians and the dispersions are found to be nearly the same for similar age groups, but in the grade groups, City B has higher means and greater dispersions than City A. Pressey (1933) commented on this situation as follows:

> The two school systems have, therefore, created a situation educationally different out of almost identical intellectual ma-

terial. Further study showed the grade differences to be due primarily to different promotion policies in the two towns. School System B retarded its children to such an extent that the average chronological age per grade was about five months above that for System A. This excessive retardation so discouraged a great many of the children that they dropped out of school as soon as possible and therefore never reached the upper grades. School System A, on the other hand, pursued a liberal promotion policy and tended to promote every child every year. As a result, the average age per grade was less than in School System B, and the children were so encouraged that more of the duller ones continued into the upper grades; even those who dropped out of school at the legal age were in a higher grade than the similar class of children in System B. These contrasting promotion policies caused System B to require a higher standard of ability for entrance to each of the grades than did System A. The higher standard in the former may appear desirable until one remembers the large number of unhappy, retarded children who never continued beyond the middle grades of elementary school. A heavy retardation will undoubtedly raise the grade averages, not only in intelligence but also in achievement, but at the expense of the best possible individual development of the children.

In making comparisons of communities and schools, one must always be alert to the differences that are caused, not by the innate ability of the pupils but by the artificial grade grouping resulting from the particular promotion policies adopted by school officials. Unless one knows the chronological age of children in the same grades in different schools or school systems, he is quite unable to make any valid comparisons as to the meaning of the test scores.

A. Accomplishment Quotients

One of the cardinal objectives listed by nearly all authorities is the development of effectiveness in learning. Since one's effectiveness depends upon both his native ability and acquired skills, it was thought that relative effectiveness would be indicated by comparing them. One method of comparison used the ratio of two test scores; one, a general intelligence test; and the other, an educational achievement test. Franzen (1922) divided the educational age by the mental age and secured an AQ or accomplishment quotient. Such quotients have nearly gone out of use, however, because of the following reasons:

1. It was shown by Rand (1925) that mental ages usually have larger dispersions than educational ages for the same group of persons.
2. Logically, the correlation between AQ's and IQ's would always be zero unless the correlation between IQ and EQ (Educational Quotient) is 1.00. This correlation is rarely or never found because the two types of tests emphasize different skills.
3. Marked difficulties have been experienced in standardizing both educational and mental ages, particularly above the average adult level.
4. The usefulness of the ratio has been questioned, for the concept of general intelligence as a unitary element has been attacked and, in the minds of many authorities, replaced by more descriptive items.
5. Intelligence quotients can rarely be taken as good indicators of native abilities.

Nevertheless, the problem of evaluating relative effectiveness is still present and still urgent. It will only be solved by longitudinal studies of development under controlled conditions. Some of these studies are described in Chapter 24.

VII. CORRELATIONAL ANALYSES

In spite of a fairly large number of reports of correlations between various achievement tests, the appraisal of independent factors is still a perplexing research problem. It was pointed out above that the various divisions of a general achievement test are seldom reliable enough for analytical work, and that the total scores are not suitable. Nevertheless, several fairly comprehensive studies have been reported and four of these will be described.

On the basis of a large number of studies of individual idiosyncrasies on achievement tests, some of which he reviewed and some of which he carried out, Kelley (1927 and

1935) came to the conclusion that there were eight essential traits or lines of cleavage in mental abilities which he called:

1. Verbal Intelligence
2. Social Intelligence
3. Activity and Mechanical Intelligence
4. Numerical Computation
5. Spatial Relationships
6. Verbal Memory
7. Special Sensory and Motor Abilities
8. Interests in Various Skills and Arts

Kelley intentionally omitted a category of general intelligence because he felt that it could be analyzed into the more independent traits which he listed, particularly the first six items. He did not claim that these were indivisible elemental patterns of behavior, but rather convenient broad groupings of patterns which had been found to be independent to a marked extent. Whether or not a general mental factor is found seems to depend at present entirely upon one's assumptions concerning its existence. If one assumes that there is such a factor, he follows Spearman's method of analysis, which nearly always yields a large general factor. If one does not wish to assume its existence, Kelley's or Thurstone's methods are followed. More research is needed to show which analysis represents the true relation of abilities most adequately.

Because English and arithmetic have long been considered basic subjects, their relationships have been studied by a number of authors. Some of these studies, which are summarized in Table 49, show a marked tendency for lower correlations in older groups. These correlations are not strictly comparable because the types of patterns measured at various ages are only roughly alike. For instance, the Arithmetic Reasoning Test requires a larger vocabulary and more complex types of calculation at the 12th year than at the 9th year. The skills leading to success at the 9th year may be inadequate or even detrimental at the 12th year.

TABLE 49

Changes with Age of Correlations between Arithmetic and Language Achievement Tests

Author	*N*	*Year or Grade*	*Average Correlation*
Thorndike (1926)	126	5th grade	.52
Schiller (1934)	186 boys	3rd, 4th grades	.63
	206 girls	3rd, 4th grades	.60
Asch (1936)	*79 boys	9 yr.	.68
	79 boys	12 yr.	.43
	*82 girls	9 yr.	.63
	82 girls	12 yr.	.30
Garett, Bryan, Perl (1935) Vocabulary and Arithmetic	306 boys	9 yr.	.52
	96 boys	12 yr.	.61
	102 boys	15 yr.	.37
	340 girls	9 yr.	.40
	100 girls	12 yr.	.55
	123 girls	15 yr.	.55
Buckingham (1937) Co-operative Algebra, and Gates Reading	105 pupils	9th grade	.38
Garett (1928)	338 men	1st yr. college	.21
Schneck (1929)	210 men	2nd yr. college	.14

* The same children retested 3 years later with the same tests.

These correlations are also not strictly comparable because we do not know how much they are affected by a narrow selection of students. College groups are often selected from the highest third of the population — a process which reduces correlations considerably from what they would have been if the total adult population had been sampled. Asch's (1936) study, however, avoids this sampling variation by using the same boys and girls. After a three-year-interval between tests, he reported that the girls showed a greater reduction in correlation than the boys. Table 50 shows the correlation coefficients for seventy-nine boys who took eight achievement tests at approximately 9 and 12, and Table 51 shows the three factor loadings secured from separate analyses of the 9- and 12-year-levels. The first factor in both levels is fairly large in all tests, but largest in the Arithmetic Reasoning and smallest in the United States Army Beta. Asch calls it a general factor. The second factor discriminates between Verbal Information and Arithmetic Calculations; the third factor is too small to be clearly identified in the 9-year-level, but in the 12-year-level it seems to represent Verbal Reasoning or deductions. Owing to the small number of tests, this analysis is far from complete and the results would have been more significant if both the 9- and 12-year results had been analyzed together. However, a tendency for test results to become more independent in the older age groups is clearly illustrated, and a tendency for the Vocabulary and Analogies Tests to become more dependent upon special factors at the later ages is also apparent. There is urgent need for research which will supplement this excellent study by using tests which do not change qualitatively with age as much as these apparently did change.

Typical of studies in which large numbers of children have been given survey tests is that of Burt (1939), who correlated scores on 13 achievement tests of 10-year-olds in several cities in England. The tests were for English Composition, Reading Comprehension, Reading Speed, Dictation, History, Geography, Science, Arithmetic Problems, Arithmetic Rules, Handwork, Drawing, Writing Quality, and Writing Speed. Using a modification of Spearman's technique, he obtained four factors from his analysis: a *General* factor which accounted for 36 per cent of the total variance; a *Verbal* factor, 9.6 per cent; a *Manual* factor, 5.8 per cent; and an *Arithmetic* factor, 2.9 per cent. These results are interesting because they are essentially the same results from a similar group measured by Burt in 1917.

The presence of other factors involving motivation and behavioral difficulties are not disclosed by the three studies just reported, but they are appraised in part by Jastak (1938). Jastak compared four groups of children: (*a*) 205 well adjusted, (*b*) 60 moderately well adjusted, (*c*) 43 moderately maladjusted, and (*d*) 41 seriously maladjusted. These ratings were made by principals and teachers. Gen-

TABLE 50

Intercorrelations of Eight Variables at the Ages of Nine and Twelve: Boys (N = 79)

Upper figures are the correlation of the nine-year-level; lower figures are the correlation of the twelve-year-level

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
1. Number	.922	.641	.420	.540	.549	.291	.512	.484
Series	.945	.627	.250	.480	.432	.326	.414	.341
2. Arithmetic		.920	.663	.668	.739	.614	.683	.491
Reasoning		.925	.456	.483	.575	.493	.433	.452
3. Vocabulary			.939	.598	.681	.651	.499	.269
			.870	.539	.709	.604	.137	.052
4. Verbal				.900	.703	.540	.550	.459
Analogies				.921	.570	.385	.119	.284
5. Sentence					.958	.712	.658	.449
Completion					.881	.734	.172	.216
6. Gates Silent Reading						—	.517	.364
						*—	.243	.206
7. Arithmetic Computation							.981	.480
							.967	.234
8. Beta (Army)								—
								*—

Average of all 9-yr. r's = .562; 12-yr. r's = .408

* Not given.

(Asch, 1936, p. 19. By permission of the Editor, *Archives of Psychology.*)

TABLE 51

Multiple Factor Analysis of Eight Variables: Boys

Variable	*I*	*II*	*III*	h^2
	NINE-YEAR-LEVEL			
Number Series	.678	− .397	− .237	.673
Arithmetic Reason	.870	− .005	− .026	.758
Vocabulary	.742	.345	− .210	.713
Analogies	.791	.002	− .101	.636
Sentence Completion	.869	.080	− .020	.762
Reading	.731	.443	.291	.816
Computation	.761	− .105	.147	.611
Beta	.579	− .326	.206	.484
ΣK^2	4.595	.597	.260	
ΣK^2	.574	.075	.033	
	TWELVE-YEAR-LEVEL			
Number Series	.675	− .348	.066	.581
Arithmetic Reason	.801	− .176	.090	.680
Vocabulary	.668	.500	− .119	.709
Analogies	.662	.148	.514	.725
Sentence Completion	.780	.394	− .066	.799
Reading	.719	.293	− .371	.741
Computation	.422	− .377	− .280	.398
Beta	.432	− .354	.023	.312
ΣK^2	3.503	.932	.512	
ΣK^2	.438	.117	.064	

(Asch, 1936, p. 23. By permission of the Editor, *Archives of Psychology.*)

eral Vocabulary Tests from the Stanford-Binet (1916) were used to yield vocabulary age. Scores of five other tests — cube construction, mazes, drawings from memory, reversing of digits, and reproducing of ideas from paragraphs read by the examiner — were combined to give a performance or alertness age. He reported that the correlation between vocabulary and performance ages for the well-adjusted *a* group was approximately .86; for the *b* and *c* groups, .74; and for the *d* group, .52. Children who were regarded as very serious problem cases frequently showed lower alertness than vocabulary scores. Children of parents who spoke a foreign language showed marked vocabulary difficulties, even at the upper school grades.

VIII. APPLICATIONS

Educational Achievement Tests have three main uses: individual diagnoses, predictions of individual success, and evaluations of the effects of instruction. There is a very wide use of tests for diagnoses of individual differences among normals, retarded and genius groups, delinquents, and those with special defects or behavior problems. Figures 17, 28, 33, and 39, and Tables 36 and 54 show diagnostic profiles. From such profiles it is possible to point out a person's strong and weak points and to suggest remedial plans. This is the work of educational and clinical psychologists, concerning which one should consult special texts.

A. Scholastic Predictions

The prediction of school success is especially important for selecting and counseling students. Among the causes of variations in making predictions, the two major sources are inconsistencies in methods of assigning school grades and large changes in the interests of pupils. With these limitations, prediction correlations of .70 are usually as high as may be expected. A few typical results are:

1. Correlations between achievement test scores in particular subject and class grades in that subject usually range from .42 to .70 (Kohn, 1938; Gates, 1922).
2. The prediction of class grades in a particular course from grades in an earlier course in the same field range from approximately .40 to .70 in the usual school or college group. Greene and Jorgensen (1936) and Williamson (1937) noted a tendency for the accuracy of scholastic predictions to decrease.
3. The prediction of algebra and geometry grades from Arithmetic Tests is usually in the neighborhood of .50, from special aptitude tests, .55, and from special aptitude and English Comprehension Tests combined, .60 (Ayres, 1934; Richardson, 1935; Orleans, 1934).
4. Prediction of general scholastic average from one year to the next is usually about .60 in large groups taking different courses (Finch and Nemzek, 1934). When all persons take nearly the same courses in the same order, correlations approximate .75.
5. Prediction of general scholastic average from a group verbal intelligence test is usually near .70 for elementary school groups, but considerably lower for high school and college groups.
6. Group intelligence tests (see Chapter 11) correlate about .80 with each other and .70 with the Stanford-Binet tests.
7. All types of correlations usually are lowered by inserting longer intervals between the two appraisals which are being compared (Kelley, 1935).
8. All types of correlations are raised by correcting for chance errors of measurement or small samples of a population. Thus Kelley (1927) showed an estimated true correlation of the National Intelligence Scale A with the Thorndike-McCall Reading Test of .95, and with the Wood-McCall Arithmetic Test of .91.

Since the actual correlations are too low for individual use, a large number of studies have been made which show that the combined scores of several tests often yield slightly better predictions than a single score. Only one study of this sort will be described.

1. PREDICTION FORMULAS. Jones and Laslett (1935) studied the prediction of average college grades from hi

school grades and the American Council on Education (A.C.E.) Psychological Test for College Freshman. The sample included 362 graduates and 138 seniors at Oregon State College. Average college success was calculated by dividing the credit hours by semester hours. Credit hours, also called quality points, were assigned as follows: 3 points for *A,* 2 for *B,* 1 for *C,* 0 for *D,* and one subtracted for each hour of *F.* (No *E*'s were used.) The high school grades were not taken at face value but reduced to a common basis according to the type of high school and the grading system in use. Mean high school scores in each of six major fields, English, mathematics, social sciences, natural sciences, foreign languages, and nonacademic subjects, were computed and correlated with the college mean. The correlations showed that no single subject-matter field predicted college success much better than the rest.

Three combinations of high school grades and A.C.E. examination scores were then correlated with college grades. One combination, called the Jones' formula, used the mean of one's high school majors and 1/100 of the A.C.E. score. This combination correlated .65 ± .02 with the college graduation grades, and it overestimated the college grades by a mean of .21 of one credit-hour point. The mean predicted mark was 1.71, and the actual mean was 1.50.

Another formula, by Dvorak, multiplied high school grades by various constants in order to give most weight to those courses which in another study were found to show the highest prediction. The constants were English, .0950; mathematics, .0157; social sciences, .2464; natural sciences, .0977; foreign languages, .1159; nonacademic subjects, .0989; and the A.C.E. Examination, .0022, plus a weight of .700. This formula allows about 16 times as much weight for social science as for mathematics work. It predicted college success .59 ± .03.

A third formula, by Taylor, combined centile of high school excellence with centile of the A.C.E. score. High school excellence was secured by counting the number of *A*'s, after first converting the grades to make them more comparable. This formula correlated with college success .64 ± .02.

A need was shown for a different formula for predicting grades of men and women, for the women made better grades in both high school and college than the men, but secured considerably lower scores on the A.C.E. tests.

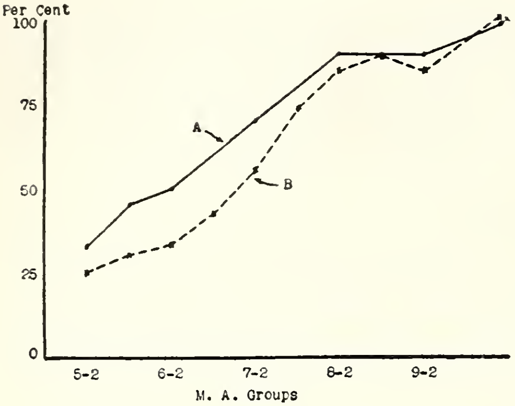
B. Optimum Age for Instruction

Typical of studies which aim to determine the more appropriate age for teaching a particular skill is that of the Committee of Seven of the Northern Illinois Conference on Supervision, reported by Washburne (1939). In 1926 this committee began investigations which have involved 30,744 children in 255 cities. Their procedure was as follows:

1. To define units of arithmetic very precisely and to devise tests to measure these units. Usually the Compass Arithmetic Test or similar tests have been used.
2. To determine the usual grade-placement of a unit by a rough survey.
3. To secure co-operation of school and teachers in the teaching of a unit at one grade lower, and at one or two grades higher than the usual grade. Standard teaching procedures have been carefully described.
4. To administer five tests: a Verbal Intelligence Test, a Pretest in Arithmetic, a Teaching Test, a Final Test, and a Retention Test, six weeks after the final test. All of these tests except the first were practically equivalent forms.
5. To discover the relationships between these tests and to suggest optimal age and grade placements. The Committee has usually felt that a unit should be taught when three-fourths of the children succeed in solving 75 per cent of the items of a retention test.

The results of one such procedure are shown in Figure 36, where the mental ages of children are plotted against per cents of success on the total test. From this figure it appears that 33 per cent success was reached by the average child at a mental age of 5 years, 1 month; 75 per cent at 7 years, 2 months; and 100 per cent at 9 years, 10 months. If a teacher desires any particular degree of mastery, the unit can be placed accordingly; or, *vice versa,* if a particular degree of mastery is found among a given group, the success of the instruction may be appraised.

The use of verbal intelligence tests in this connection might be challenged on the ground that readiness for a particular arithmetic unit is a special ability which may not depend very much upon the skills needed for success on the intelligence tests. Much research is still needed to show the relationship between various predictions of arithmetical success. The Committee has, however, published fairly detailed and useful statements of the skills which have been mastered by particular mental age groups, one of which is given below:



A. Per Cent of Items Passed by MA Group
B. Per Cent of Group Retaining 75 Per Cent or More Items

FIGURE 36

Growth of Success on Hard Addition

(After Washburne, 1939, Figure 1. By permission of the National Society for the Study of Education.)

Mental Age 7–8

The addition facts with sums of 10 and under are well learned at this level, and there is little gain in further postponement. The harder addition facts and the easy subtraction facts can be successfully learned at this age, but there is a definite gain in postponing them to the next level. The desirability of systematic drill in these facts at this level is open to question, in spite of the fact that it produces satisfactory results. Many persons feel and there is some evidence to justify the feeling, that the informal experiences and activities of mental level 6–7 should be continued and extended at this level and that systematic drill of all sorts should be postponed to the next one.

Simple comparisons of length, height, thickness, width, and the like, including the recognition that one object is two, three, or four times as high, wide, long, thick, or deep as another, are well learned. Children can also readily learn to measure lines in even inches, and, with more difficulty, to draw lines an even number of inches long. They can learn how many inches there are in a foot and in two feet.

Children can learn to read the clock on the even hour, to distinguish between morning and afternoon, and to understand the symbols *a.m.* and *p.m.*

C. Vocational Predictions

One of the most elaborate studies of prediction of vocational success from school success is that of Thorndike (1934) and his associates. They reported the relationships between three sets of appraisals: school record of the usual sort, tests during the eighth grade, and records of work ten years later.

Two groups of students in New York City were studied. One group of 271 boys and 203 girls was selected from schools which served families of low economic status. The other group of 826 boys and 925 girls represented the entire city population fairly well except for the most retarded pupils who are excluded from school. By diligent work, complete records were obtained for approximately 78 per cent of boys and 82 per cent of girls. The students whose records were incomplete because they moved away were not significantly different from the others at age 14. Students who could not be located or who refused to co-operate were slightly inferior to the others at age 14 in school progress, intelligence test scores, and scholarship.

The school records which were secured together with their retest reliabilities are shown below:

1. Age in eighth grade, third month (.99)
2. Progress during school attendance (.92)

3. Conduct (.99)
4. Scholarship (.96)
5. Attendance (.99)
6. Age at leaving school (.97)
7. Grade at leaving school (.99)

The tests used in the eighth grade and approximate reliabilities for an age group were:

8. Clerical intelligence (Toops, .85)
9. Clerical activities (.80)
10. Stenquist Assembly Test (.40)
11. IER Assembly Test for girls (.70)
12. Arithmetic problems (above .80)
13. Language tests (above .80)
14. A combination of tests 12 and 13 called intelligence (.85)
15. Average annual earnings at ages 20 to 22 (.90)
16. Average level of jobs (.70)
17. Average liking for jobs (reliabilities not calculated, thought to be about .70)
18. Per cent of times employed (.90)
19. Number of changes of employer (thought to be near .99)

These retest reliabilities of school records and vocational records are very high. The variations were largely due to clerical errors and slight shifts in standards. The reliabilities of the tests and estimates used are considerably lower, although they are as high as is generally found for material of this sort. The test reliabilities are not high enough for good individual prediction, but they are high enough to show group trends.

All of these items were intercorrelated in order to show their predictive value. The results may be summarized as follows:

1. The age of leaving the eighth grade, scholarship, and intelligence test scores predict fairly well the grade which will be reached at later ages. Indirectly, this finding is of vocational significance because the grade which can be completed indicated the level of college or professional work that may be attempted.
2. Among the 223 men and 247 women who did clerical work for at least nine-tenths of the time at ages 20 to 22, it was

found that clerical earnings correlated with the earlier tests of clerical intelligence, .26, with clerical activities, .22, and with scholarship, intelligence, and mechanical assembly to smaller degrees. Conduct and attendance showed zero relationships with clerical earnings. The highest correlation obtainable was approximately .30 for boys and .40 for girls. The correlations between level of jobs and interest in work and the school records and test were all nearly zero.

3. In the case of the 210 men and 155 women who worked nine-tenths of the time at mechanical work from ages 18 to 20, none of the school records or tests showed significant correlations with earnings, liking for work, and interest. This situation was also typical of the 299 men and 76 women who had combinations of mechanical and clerical work of other varieties.
4. The higher a pupil's score in clerical and intelligence tests and in scholarship, the more likely was he to do better at clerical than mechanical work and *vice versa,* but the likelihood was not very large.
5. Of the eighth grade boys, 20 per cent attended college for at least one semester; and of the girls, 12 per cent.
6. About 2 per cent in the group became criminals. These were inferior to the group average at age 14 in all respects.
7. The frequency of change of employer had slight and probably insignificant relationships to earnings, level of occupation, and interest.
8. The annual earning in white collar jobs was greater than in mechanical jobs among women, but not among men.
9. There was much evidence that employers did not select employees on the basis of ability alone. Had this been the case, the prediction from eighth grade records would have been materially higher.
10. The prediction of later success, if the study is carried on further, will probably be greater because many persons had not yet shown what they could do vocationally.

This study points to the need of more careful instruments for appraising abilities at early ages and factors in success at later ages. It indicates that when more precise measurements are available and when allowances are made for such disconcerting factors as health, and racial prejudice, prediction of vocational success will be of considerable accuracy.

10

BINET–TYPE SCALES

I. CHARACTERISTICS

The individual tests that have been, and still are, applied most widely, both in this country and abroad, may in most cases be classified as of the Binet type. No accurate estimate of the number of such tests given annually is at hand, but it can be safely said that they are applied almost universally in studies of behavior difficulties and delinquency in children, and to a lesser extent in studies of adults.

The characteristics of Binet-type tests are:

1. The use of a wide variety of tasks which are designed to distinguish between bright and dull persons, but which are not designed to measure analytically any particular skills or body of information.
2. The administration of the tests individually, using oral directions. The order of presentation may be varied somewhat at the discretion of the examiner to secure a good sample of the subject's ability.
3. The use of a total score rather than a profile of separate scores. The score is usually given in terms of mental age and intelligence quotient, or corresponding indices.

4. An arrangement of test items according to levels of difficulty called mental age levels, or else an arrangement simply according to relative difficulty.

II. EARLY SCALES

Alfred Binet, born in 1857, in Nice, France, studied medicine under Charcot and Féré, and with Beaunis founded the Psychological Laboratory at the Sorbonne in Paris in 1889. He was an energetic worker, and during his professional life he conducted a very large number of technical researches. When in 1895 he became particularly interested in mental measurement, he and a student named Henri evaluated the tests which were then available. They decided that there was need for fewer tests of sensory processes and for more tests of complex thought processes. In 1898 Binet described current tests, most of which were not of his own invention, but which he thought would indicate accuracy of judgment and general mental ability. Specifically, he mentioned the following tests: mental calculation, drawing a square, reconstruction of disarranged sentences, comprehension of abstract passages, questions of moral or social propriety, immediate memory for numbers and for objects, and imitation of paper folding. It is interesting to note that nearly all of these tests with slight variations are used in practically all of the more recent Binet-type tests and also in many group tests of intelligence. Binet also emphasized the idea that mental measurement should result in a rating of persons with reference to one another rather than an absolute rating of ability.

In 1900 Binet published "Attention and Adaptation," which reported a study of the differences between two groups of children: five who had been classified as bright by their teachers and principal; and six, as dull. The children of both groups averaged approximately 11 years of age. The two groups showed no large differences on tests of simple or of choice reaction time, perception of small variations in metronome speed, immediate memory for words, and speed of counting dots. He did find large differences on tests of tactile sensibility on the back of the hand; copying letters, words, and designs; cancelling letters from a printed page; and addition. He concluded that the bright showed a quicker and more accurate perception and solution of difficulties, and he defined attention as mental adaptation to a new situation. He was convinced that speed of routine acts had no relation to intelligence.

After various studies of cephalic indices and two-point thresholds on the skin, he again turned to mental tests and reported in 1902, a large number of tests of free association of verbal processes on his two daughters: Marguerite, age fourteen and one-half, and Armande, age thirteen. No quantitative results were given, but Marguerite was clearly shown to be precise, practical minded, well oriented in space but not in time, and little given to reverie. Armande was more imaginative, vague, detached, delighted in fantasies, and had a tendency to verbalism.

In 1904 he worked with Simon in comparing feebleminded and normal children, and they published in 1905 a list of thirty tests which were arranged in order of difficulty. These items, shown in Table 52, were thought to depend more on broad cultural experiences than on specific academic training.

On the basis of testing approximately fifty normal children and a larger number of mentally retarded children called *aments,* tentative norms were given. The first five tests were found to be passed by idiots and 2-year-olds. The ninth test was the upper limit for normal 3-year-olds, the fourteenth test for 5-year-olds, and the fifteenth for adult imbeciles. The sixteenth item separated 5- from 7-year-olds effectively. Morons fell below the 12-year-level in the test, and were more clearly distinguished from normal children by verbal than by nonverbal tests. The discrimination be-

TABLE 52

The 1905 Binet Scale

1. Visual co-ordination of head and eyes
2. Grasp a cube placed on the palm
3. Grasp a cube held in line of vision
4. Make a choice between pieces of wood and chocolate
5. Unwrap chocolate from paper
6. Obey simple orders; imitate gestures
7. Touch head, nose, ear, cap, key, and string
8. Find objects which the experimenter names in a picture
9. Name objects pointed out in a picture
10. Tell which of two lines is the longer
11. Immediate memory for three digits
12. Tell which of two weights is the heavier
13. Suggestibility:
    a. Find object which is not among those presented, as in No. 8
    b. Point to *patapoum* and *mitchevo* (nonsense words) in the picture No. 8
    c. Tell which of two equal lines is the longer
14. Give definitions of *house*, *horse*, *fork*, and *mama*
15. Immediate memory for sentences of fifteen words
16. Give differences between: paper and cardboard; fly and butterfly; wood and glass
17. Immediate memory for thirteen pictures of familiar objects
18. Immediate memory for two designs, exposed ten seconds
19. Immediate memory for list of digits, three, four, or five in the series
20. Give similarities between: blood and wild poppy; fly, butterfly, and flea; newspaper, label, and picture
21. Just noticeable differences in length of lines
22. Arrange three, six, nine, twelve, and fifteen gram weights in order
23. Find which weight has been removed from No. 22
24. Find rhymes
25. Complete simple sentence by adding one word (after Ebbinghaus)
26. Construct sentence containing *Paris*, *gutter*, *fortune*
27. Knowledge of what is the best thing to do in twenty-five situations of graded difficulty
28. Reverse clock hands at 3 : 57, at 5 : 40, and tell the time it would be
29. Draw results of folding a piece of paper into quarters and cutting the once folded edge
30. Distinguish between liking and respecting, between being sad and bored

(Peterson, 1925, p. 172. By permission of the World Book Co.)

tween higher age groups was made on the basis of type of answer. The authors realized keenly the need of more precise standards and more cases.

Working independently in Worcester, Massachusetts, Terman (1906) published a comparison of seven superior and seven dull boys on tests of ingenuity, logical processes, mathematics, mastery of language, interpretation of fables, learning to play chess, immediate memory for numbers, for forms, and for stories, puzzle solution, and motor ability. A number of these tests had been described in earlier works, but most of them were originated by Terman. He found that the dull boys, who were older and stronger, excelled the bright on motor tests, but fell behind them on the mental tests. Many of these tests were standardized later.

At the University of Rome, also working somewhat independently, Sante de Sanctis (1906) designed six tests to be used for classification of the feeble-minded. His battery emphasized what are now known as performance tests, in which one had to select and manipulate various colored balls and cubes and plane figures.

At the Vineland Training School in New Jersey, Goddard (1913) had collected and devised twenty-five tests to be used in the diagnosis of feeble-mindedness before he saw Binet's work. His tests included the Seguin form board, see Chapter 12.

In 1908, Binet incorporated the mental level concept into a new scale of fifty-nine items. Three to eight items were selected to represent each age level from 3 to 13 years. Each test was assigned a certain number of months' credit, and a person's score was the total years and months of mental age which he had earned. The scale was arranged so that the average child of a particular chronological age would have a corresponding mental age.

This work was eagerly read. Translations of the 1908 scale were applied in Italy by Ferrari (1908), and by Treves and Soffiotti (1909); in Switzerland by Descœudres (1911); in Germany by Bobertag (1912); in America by Goddard (1911), Kuhlmann (1911), Terman and Childs (1912); and in England by Johnson (1911). All of these investigators reported that Binet's tests resulted in too large a percentage of children rated as superior below the age of 10 years, and too small a percentage in the higher age levels.

However, Stern (1914) considered that this agreement of findings from six different countries using different languages and testing children from different environments was very strong evidence that the tests measured the "general developmental conditions of intelligence — and not mere fragments of knowledge and attainments acquired by chance."

In Germany, Stern (1900) published a book on differential psychology and Meumann (1905) made an extensive review of test literature which indicated notable advances in Germany before Binet's first scale became widely used. Meumann pointed out the need of an indication of rate of development, because the older feeble-minded children were more retarded in mental age than the younger. To meet this need, Kuhlmann (1912), Stern (1914), and Yerkes, Bridges, and Hardwick (1915), apparently working independently, suggested indices to be found by dividing an individual's achievement by the age group average. Terman (1916) divided the mental age (MA) by the chronological age (CA) and called the resulting index an intelligence quotient (IQ).

Terman became so much interested in the 1908 Binet scale that he applied a translation of it to four hundred nonselected children. He was then convinced that the scale was practical in spite of imperfections in sampling and scaling techniques. He set about to correct some of these defects. The idea, which Binet and others had often repeated, that an adequate measure of intelligence can only be secured from a variety of tests was accepted because the pooled results seemed to distinguish between bright and dull better than any one single type of test.

Binet's last scale appeared in 1911, the year in which he died. Several reading and writing tests which he believed depended too much on special training were omitted. He relocated many of the items, particularly in the higher age groups. These changes resulted in a scale which had five items at each age from 3 years to adult, omitting the 11-, the 13-, and the 15-year-levels.

III. TERMAN'S REVISIONS

In 1916 Terman published a revision of the 1911 Binet scale upon which he and his associates had been at work for four years. The original items were restated to remove ambiguities in administration and scoring. The scale included tests for ages from three to adults of superior ability. It was rapidly adopted as the standard individual test of intelligence by most of the schools and clinics in America.

Terman's 1916 scale was replaced after twenty years of very wide application by a revision published in 1937 by Terman and Merrill. Most of the items of the 1916 scale were retained and the scale was extended down to the 2-year-level, and up to a 22-year-level. The 1937 revision provides two equivalent forms, called Forms L and M (see Table 53). Both forms contain tests for each six-month-level from 2 to 5 years, and for each twelve-month level from 5 to 14 years, and for four higher levels. The four highest levels are called the Average Adult, Superior Adult I, Superior Adult II, and Superior Adult III. An inspection of Table 53 gives a rough idea of the types of processes used in these scales. At the earlier ages, objects, pictures, and parts of the body are more often included in the test situation than at later age levels. Tests of immediate memory spans for words and numbers are found throughout the scale. At the higher levels are found more items which seem to depend upon abstract verbal and numerical reasoning. The reasoning elements at any age level seem to the writer to contribute less to the average score than recall of learned factual material. A large box of standard materials is used at age levels from two to six years. Some of these materials are pictured in Figure 37. Cards with standard pictures and other printed material are also used at later ages.

TABLE 53

Revised Stanford-Binet Tests of Intelligence

Partial Outline of Form L

YEAR II

*(6 tests count 1 month each, or 4 *tests, 1½ months each.)*

1. * Place three small blocks into similar holes in a board.
2. Point to toys when their names are given.
3. * Point to parts of a large paper doll when parts are named.
4. Build a four-cube tower after demonstration.
5. * Name common objects from separate pictures.
6. * Use a two-word sentence spontaneously; See kitty.

Alternate: Obey simple commands to manipulate small toys.

YEAR III–6

*(6 tests count 1 month each, or 4 * tests count 1½ months each.)*

1. * Obey simple commands to manipulate small toys.
2. * Name common objects from separate pictures.
3. Point to the longer of two sticks.
4. Name at least three objects shown in one picture.
5. * Point to objects to indicate use: Show me which one we drink out of.
6. * Tell what to do in common situations.

Alternate: Draw a cross with a pencil after demonstration.

YEAR VI

*(6 tests count 2 months each, or 4 * tests count 3 months each.)*

1. * Define five words orally by description, use, or classification.
2. * Make a simple bead-chain pattern from memory after a demonstration.
3. Tell what part is missing from four pictured objects.
4. * Select certain numbers of blocks from a pile.
5. * Point to one of five pictured objects which is different from the rest.
6. Draw a pencil line through a sample maze to make the shortest path.

YEAR X

*(6 tests count 2 months each, or 4 * tests count 3 months each.)*

1. * Define eleven words orally.
2. Explain why the pictured actions of a person are foolish.
3. * Read a passage of 48 words, then recall from memory a considerable portion of it.
4. * Give two reasons to support an oral statement.
5. * Name as many disconnected words as possible in one minute.
6. Repeat six digits after one oral presentation.

AVERAGE ADULT

*(8 tests count 2 months each, or 4 * tests count 4 months each.)*

1. * Define twenty words orally.
2. * Transcribe a short message in a code which is exposed.
3. * Give differences between two abstract words.
4. Read short arithmetical problems and give answer without using paper and pencil.
5. Tell what proverbs mean in own language.
6. * Give oral solution of a practical mechanical problem which is presented orally.
7. After one oral presentation, repeat a 24-syllable sentence without error.
8. Tell in what way verbal opposites are alike.

SUPERIOR ADULT

*(6 tests count 6 months each, or 4 * tests count 9 months each.)*

1. * Define 30 words orally.
2. Read aloud a problem concerning direction and distance travelled, and give answers without using paper and pencil.
3. * Give opposites of words by analogy.
4. * Watch examiner fold and cut a piece of paper, then make a pencil drawing to show how the paper would look if it were unfolded.
5. * Read silently while the examiner reads aloud a simple geometric progression problem, then give answers without using paper and pencil.
6. Repeat 9 digits after one oral presentation.

* Tests included in shorter form.

(Arranged from Terman and Merrill, 1937. By permission of the Houghton Mifflin Co.)

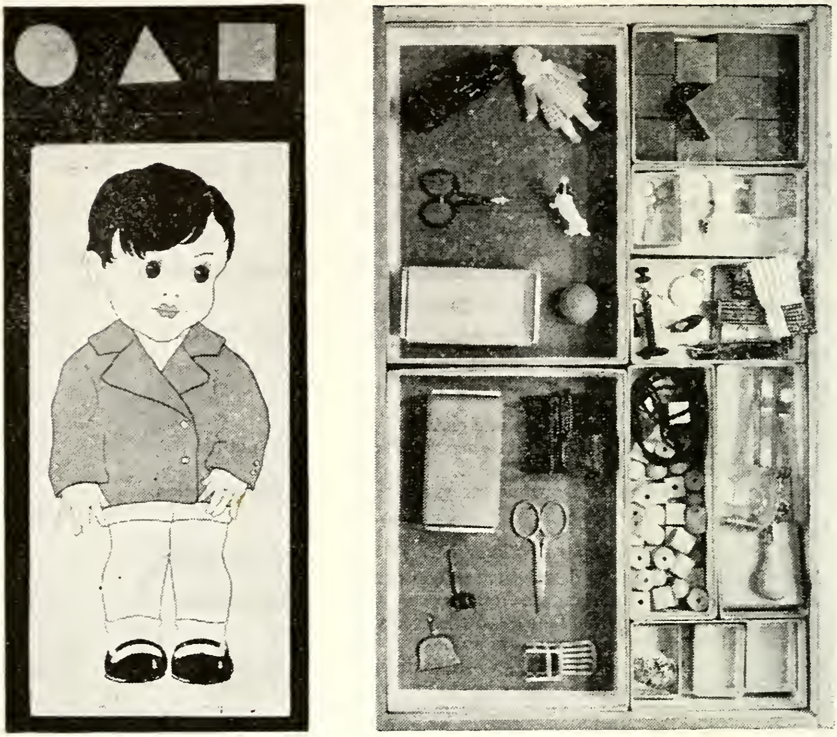


FIGURE 37. **Stanford-Binet Test Material**

(Terman and Merrill, 1937. By permission of the Houghton Mifflin Co.)

No quantitative analysis of the mental processes needed for success has yet appeared, but the authors have promised a volume which will fill this need. They do, however, discuss the tests qualitatively. Thus, page 235 evaluates Verbal Absurdities at the VIII-year-level as follows:

> The detection of absurdities has again shown itself to be one of the most valid and serviceable of our tests. No other test in the scale with the exception of vocabulary yields more consistently high correlations with total score. These range from .72 to .75 for a single age group. The test appears to be little influenced by schooling or by differences in social status.
>
> The purpose of the test is to discover whether the subject can point out the intellectually irreconcilable elements of the situation presented. The only difficulty is in judging whether the response shows that the subject has seen the incongruity. The

> child who has seen the point instantly often indicates that fact by repeating the critical phrase, and the dull, uncomprehending child may just try to say over what you have said. . . .

In discussing the Memory for Designs Test in Year IX (p. 248), the authors write:

> The figures are, of course, perceived as meaningful wholes; that is, the lines of the figures constitute designs and, in so far as they are recalled, are recalled as related. Whatever may be the processes involved — attention, visual memory, kinesthesis — it is very certain that they differ from one individual to another and that the dependence, for instance, on visual cues is more marked in some children than in others. It is often possible to note the utilization of kinesthetic cues as the child practices the designs with pencil in air during the ten second exposure interval.
>
> For *half credit* all of the elements must be present, but inaccuracies due to omission or addition of details or to irregularities in size and shape of the figures are overlooked. The samples on pages 250–251 indicate the standard for plus, half credit, and minus.

IV. KUHLMANN'S REVISIONS

Kuhlmann (1922) brought out a revision of the 1911 Binet upon which he and some of his staff had been working for seven years. It was a remarkably thorough piece of work, which extended the scale downward to the 3-month-level and up to a mental age of fifteen years. He included 129 items, counting a test once for each age group in which it is used. Of the original Binet tests thirty-seven were retained, but a number of these were modified or shifted to other age levels. The number of tests in each age group above 2 years was increased to eight. Credit was allowed for both speed and accuracy on many items.

Kuhlmann's (1939) *Tests of Mental Development* is the latest work of the staff of the Division of Examination and Classification of the Minnesota State Department of Public Instruction. About 3,000 nonselected public school and preschool children were used to evaluate one hundred and twenty-one tests which had been chosen from various sources, particularly Kuhlmann's earlier test and Gesell's (1928) studies of Infancy and Human Growth. Eighty-nine regular tests and nineteen supplementary tests were finally included in the scale on the basis of four criteria:

1. Large increases in raw scores from age to age, in those tests which had several elements
2. Large increases in per cents of children in successive age levels who pass a test (in those tests which were either passed or failed)
3. Variability in raw scores in a single age group (A wide distribution of scores in one test was considered desirable.)
4. Correlations of single test scores with total scores (High correlations were preferred.)

The scores below the 8-year-level, that is, on the easiest sixty-three tests, were the number of items correct. On harder tests the score used was speed-times-accuracy. Speed was usually calculated by dividing the right-minus-wrong elements by the time in seconds. The tests were usually limited to one or two minutes. Accuracy was calculated by dividing the number of correct elements by the number attempted. Multiplying speed by accuracy penalized the inaccurate worker. This penalty was imposed by Kuhlmann when he found fairly large negative correlations between speed and accuracy, although both were related to intelligence and age.

An innovation in this test battery is the use of a point scale designed by Heinis to give equal credits to equal units of growth. The Heinis growth curve is shown among others on page 673 and discussed in Chapter 21. Each score on each test is changed by the use of conversion tables to *mental units,* called MU points. A person's total score, the sum of these points, can be changed to his mental age by another table. IQ's and *Per cents of the Average* (PA)[1] may be obtained from an 82-page table. The PA is preferred by Kuhlmann since it has proved to be more constant than the IQ's in retests over a period of years. He also found that age groups had more constant standard deviations for PA than for IQ scores. The standard deviation for PA's was eight points, approximately one-half that of the IQ's. Kuhlmann believed that this smaller range of the PA's is a better representation of actual individual differences in intellect than was the larger range of IQ's.

---

[1] The PA is an index secured by dividing a person's MU points by the average MU points for his age group.

V. OTHER REVISIONS

In 1915 and in 1923 Yerkes and others published a revision of Binet's work which did not assign tests to mental age levels, but allowed various points credit for each item. The total points could be converted into a corresponding mental age by reference to a table. Point scales are well illustrated by the work of Baker and Leland (1935), who published a battery of nineteen tests, each of which contained from 14 to 360 points. Table 54 shows the names of the tests, the mean scores for each age group, and the profile of a 10-year-old boy. This battery includes the main types of items which are included in the Stanford-Binet Forms, and in addition a hand-and-eye co-ordination test. It needs a little longer time for administration than the Stanford-Binet, but it has the following advantages:

1. It includes more items of each type, thus insuring a better appraisal of the individual
2. It allows a profile of results which shows a person's strong and weak points. This analysis is particularly valuable in remedial work and counseling

In criticism of this scale it should be mentioned that the standardization is not so complete as is desirable, and that the profile does not indicate traits which are known to be independent or unitary. However, profiles such as this

TABLE 54

Detroit Tests of Learning Aptitude

Age Norms

Verbal	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*
1. Picture Absurdities	1	3	5	7	9	11	14	(16)									
2. Verbal Absurdities					6	9	12	15	(18)	22	26	30	34	37	39		
8. Social Adjustment		1	14	24	(31)	36	40	44	48	52	56	60					
15. Social Adjustment	1	6	14	21	24	(26)	30	34	38	42	44	46	49	52	55	57	60
10. Orientation	3	7	13	18	22	26	30	33	36	38	(40)	41					
4. Verbal Opposites				6	13	19	26	34	(11)	46	50	54	58	63	69	76	86
19. Verbal Like–Different					5	8	(11)	15	19	23	27	31	35	39	45	56	76
11. Free Association			8	12	16	(20)	24	28	32	36	40	44					
Immediate Memory Spans																	
6. Auditor Words Unrelated	2	2	3	3	3	4	4	4	5	5	(5)	6	6	7	7	8	8
13. Auditor Words Related	2	10	13	15	17	18	20	21	22	(23)	24	25	26	27	28	30	32
16. Visual Letters				3	3	4	(4)	5	5	6	6	6	7	7	7		
9. Visual Objects	2	2	2	3	3	4	(4)	5	5	5	5	6	6	7	7	8	8
7. Oral Command	1	4	6	10	14	18											
18. Oral Directions				1	3	4	5	6	(7)	8	9	10	11	13	13	15	17
Nonverbal Answers																	
12. Visual Design	1	3	5	6	9	11	14	18	22	24	26	(28)	30	31	32		
3. Picture Opposites	1	4	7	9	11	13	15										
17. Broken Pictures				4	8	12	16	20	24	(28)	32	36	40	44	48		
Number																	
14. Calculations	1	5	9	13	24	41	48	52	(55)	57							

(Arranged from Baker and Leland, 1935. By permission of the Public School Publishing Co.)

probably represent fairly independent behavior patterns and they are often more useful than a single mental age score. From analysis of such profiles knowledge of independent variables will eventually come.

It would be unwise to include a more detailed description of test administration and scoring here, since the value of a test depends somewhat upon its novelty. Even with such precautions as this, coaching on tests does take place and sometimes it becomes a source of serious error in individual cases.

These brief descriptions of the scales do not give a clear picture of the actual construction work. Some knowledge of the statistical work which precedes the publication of the final form of a test seems essential for understanding and interpreting results.

VI. CONSTRUCTION OF BINET–TYPE TESTS

There are two main procedures necessary in the construction of any mental test. One is the selection of items and the other is the scaling of items.

A. Selection of Items

The principal criteria used in selecting an item for Binet-type scales are described below.

1. Discrimination between Bright and Dull. In spite of the mass of data available, little has been published lately to show which items do really discriminate best between normal and defective groups of the same chronological age. Burt (1922) and Merrill (1924) have confirmed the earlier work by showing large mean differences in scores of groups selected on the basis of school progress on the usual tests of facility in the use of language and number symbols, and smaller mean differences on motor and performance tests.

2. Correlation between Items and Total Score. An illustration of the selection of items to correlate with the

total score is seen in Table 55 A, which compares the per cents of persons of the same chronological age in three IQ groups who passed each item listed. This shows that the first item in Year VI (distinguishing right from left parts of the body), was passed by 43 per cent of the 6-year-olds whose IQ's were below 96, by 70 per cent of the 6-year-olds whose IQ's ranged from 96 to 105 inclusive, and by 89 per cent of the 6-year-olds whose IQ's were above 105. A glance at this table shows that there are fairly large differences among these IQ groups on nearly all items. Items which did not show as large differences were eliminated by Terman. Selection of items in this fashion undoubtedly increases the internal consistency of a test.

In this table, the indications of brightness are IQ's which were determined from the test items themselves. Hence the differences between IQ groups are doubtless somewhat larger than would be found among groups which were selected for relative brightness by some other method, such as teacher's ratings.

Terman and Merrill (1937) published only a few correlations between item and total scores. The highest are those for vocabulary which, for various age groups, range from .65 to .91, median .81. The lowest correlations reported are for block counting, .43; counting taps, .50; motor co-ordination, .46; and picture absurdities, .56.

Kuhlmann (1939) published correlations between total scores and separate items for age groups of approximately one hundred and fifty pupils. The median was approximately .45. There was a marked tendency for the harder tasks to show smaller correlations than the easier. Thus, the tests used below the age of 3 years had a median correlation with the total score of approximately .91; from 3 to 5½ years, of .50; and above 5½ years, of .39. This finding may indicate a much greater internal consistency at the lower ages or it may be due to other factors, such as the length of the tests and the range of scores in a group.

TABLE 55

Relative Difficulty of Items for Normal, Retarded, and Superior Subjects

Per Cents of Single-Year Groups

			A			B		
			CA CONSTANT IQ's			MA CONSTANT IQ's		
			Below 96	*96–105*	*Above 105*	*Below 70*	*90–110*	*Above 140*
YEAR VI	1.	Right and left	43	70	89	63	69	75
	2.	Mutilated pictures	32	66	86	88	63	25
	3.	Thirteen pennies	40	77	96	88	81	50
	4.	Comprehension (2)	50	68	85	75	69	100
	5.	Four coins	54	79	83	94	92	0
	6.	16–19 syllables	57	65	78	33	42	67
	Al.*	A.M. and P.M.	64	82	86	—	—	—
YEAR VII	1.	Fingers	53	68	85	91	85	50
	2.	Description of picture	48	52	80	67	70	83
	3.	Five digits	62	74	80	61	64	83
	4.	Bow knot	43	71	78	55	71	33
	5.	Differences	48	74	95	45	55	33
	6.	Diamond	38	58	82	77	64	66
	Al.	Days of work	33	62	85	—	—	—
	Al.	Three digits reversed	39	55	75	—	—	—
YEAR VIII	1.	Ball and Field	48	60	69	61	30	47
	2.	20–1, count backward	35	55	83	67	69	40
	3.	Comprehension (3)	52	80	80	76	78	100
	4.	Similarities	44	57	83	28	46	73
	5.	Superior definitions	44	60	80	61	81	87
	6.	Vocabulary (20)	26	57	74	45	43	40
	Al.	Six coins	53	64	71	—	—	—
	Al.	Dictation	65	90	100	—	—	—
YEAR IX	1.	Date	48	68	88	71	63	26
	2.	Weights	37	58	82	47	47	91
	3.	Change	39	60	73	83	67	31
	4.	Four digits reversed	45	62	88	58	67	54
	5.	Three words	49	67	91	83	65	68
	6.	Rhymes	39	67	88	59	58	91
	Al.	Months	39	55	85	—	—	—
	Al.	Stamps	40	76	91	—	—	—
YEAR X	1.	Vocabulary (30)	20	60	80	39	56	35
	2.	Absurdities	35	60	84	83	79	71
	3.	Designs	45	57	70	71	65	38
	4.	Reading and repeating	45	63	84	80	86	80
	5.	Comprehension (4)	25	64	76	67	56	57
	6.	Sixty words	30	70	76	61	71	81
	Al.	Six digits	55	67	81	—	—	—
	Al.	20–22 syllables	45	57	79	—	—	—
	Al.	Healy's form board	55	67	76	—	—	—

* Alternate.

(A. from Terman *et al.*, 1917. By permission of the Editor, *Educational Psychology Monographs;* B. from Merrill, 1924. By permission of the Editor, *Comparative Psychology Monographs.*)

The results of both Kuhlmann and Terman lead one to suspect that there are a number of highly independent factors sampled by these tests — a hypothesis discussed in Chapter 11.

3. Discrimination among Adjacent Age Groups. An illustration of the selection of items which distinguish between adjacent chronological age groups is seen in Table 56, which shows the per cents of persons who passed various items. From this table a growth curve for each item may be prepared, as in Figure 38. The most desirable item is one which shows a rapid period of growth both before and after the age which is to be measured. Such items will distinguish between adjacent age groups better than items which show very little increase during the ages to be compared. The items shown in Table 56 were selected in part because they had fairly similar and regular growth curves. Other items have been found, as in Figure 34, page 237, which do not show such similar and regular growth curves. The shapes of such curves depend undoubtedly upon both native and environ-

TABLE 56

Per Cent of Age Groups Passing Items

Test	*Age*											
	3	4	5	6	7	8	9	10	11	12	13	14
Age XI												
Absurdities	0	0	0	0	5	24	29	49	70	79	96	97
Difficult Questions .	0	0	0	1	6	13	28	48	64	76	92	95
Sixty Words	0	0	0	4	7	21	27	43	60	74	88	92
Seven Numbers . .	0	0	0	2	5	18	25	44	59	68	80	92
Sentence Building (1)	0	0	0	0	3	16	20	43	58	68	81	89

(After Burt, 1922, Table III.)

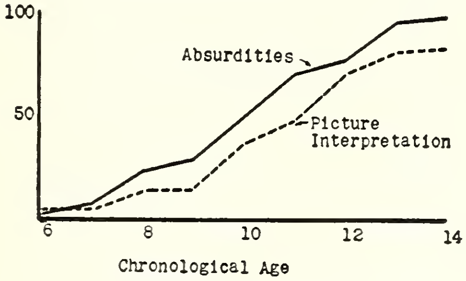


FIGURE 38

Per Cents of Age Groups Passing Binet Items

(After Burt, 1922, Table III.)

mental factors. A special training period often results in a rapid rise in a growth curve. Training is therefore an important factor in determining both the difficulty of an item for a particular age group and the resulting MA scores for persons in that group.

4. OTHER CRITERIA. In the 1937 Stanford Revision, the items were selected after six successive revisions which involved shifting items from one form to another, and sometimes modifying the scoring to make an item harder or easier. The following features were considered to be desirable in the final selection of items:

1. Ease of scoring
2. Short time requirements
3. Interest to subjects
4. Balancing or elimination of sex differences
5. Equal means and standard deviation of IQ's at various age levels

In attaining these features, 30,000 cards were used for the mechanical tabulation of the results of the testing of the 3,184 native-born white persons in the standardization group. The authors write (Terman and Merrill, 1937, p. 22):

> By means of the Hollerith sorter it was then possible to plot for each test the curve showing per cent of subjects passing in successive ages throughout the range, also the curve of percent passing by successive intervals of composite total score on the two forms. This was done for the sexes separately as a basis for eliminating tests which were relatively less fair to one sex than the other. It was possible also to compare the scores on the form given first with those on the form given second, to study the effect of practice, and to allow for it in the computation of the composite IQ's. The correlation of each test with composite total score (equivalent to correlation with mental age) was computed separately for each test, thus providing a basis for the elimination of the least valid tests. One important use of the Hollerith data was in connection with the balancing of the two scales; it was important that at each level the two scales should be as nearly alike as possible with respect to the relative difficulty of the tests located at that level and with respect to their correlation with total score.

These criteria of selecting items for a Binet-type test have often been challenged for the lack of analysis of the psychological factors involved. Brightness and age status are probably complex resultants of a number of forces which may be independent. Methods of analyzing such forces are discussed in Chapter 7. For the present, let us assume that items which seem most appropriate have been selected, and consider the next step which is the scaling.

B. Scaling of Items

Mental age is usually defined as a score on a test which represents the average of a narrow age group. For instance, the average score of a group of children just 6 years old is taken as the 6-year-standard, and persons who make this score are said to have a mental age of 6.

There are two methods of changing test scores into mental age scores. One, called point scaling, assigns points of credit to all the items in a test, sums the individual credits, and then finds the average scores of various age groups. This method results in a conversion table such as seen in Tables 10 and 54. Point scaling is easy to apply since it only requires that the items be roughly placed in order of difficulty so that the subject is presented with all the tasks on which he is likely to succeed. This method has been used by Yerkes *et al.* (1915), Baker and Leland (1935), and Kuhlmann (1939). The other method, called *age-level scaling,* assigns test items to particular age levels and allows a number of months of credit for each item. This method, used by Terman, is hard to apply because it is difficult to find items which fit exactly into a particular age level. The construction of age level scales is discussed below.

1. Assignment to an Age Level. The process of assigning an item to a particular age level is rather intricate. First, the items are administered to a fairly large group of persons of various ages. Then the per cents of persons who pass the item are recorded, as in Table 56. An inspection of this

table will show approximately at what age 50 per cent of a group pass the item, and how well any item discriminates between age groups. Here an item in the XI-year-level, *Absurdities,* shows at age 9, 29 per cent passing; at age 10, 49 per cent passing; at age 11, 70 per cent passing; and at age 12, 79 per cent passing. One is justified in concluding that it does discriminate fairly well among 9-, 10-, and 11-year groups, but not between 11- and 12-year groups.

Table 56 also shows that the 11-year group ranged from 58 per cent to 70 per cent passing among the five items in this age level. If all persons in the 11-year group were just 11 years old, items would be needed at this level which were passed by just 50 per cent of the group in order to give the average person an MA of 11 years. This 11-year-old group, however, included children from 11 years to 11 years and 11 months of age, average 11 years 6 months. Hence, if the item is to represent just an 11-year level, the per cent passing must be higher than 50 per cent.

In the 1916 Terman scale, the items assigned to each age level had various mean per cents passing as shown in Table 57. The smaller per cents passing shown in the higher age levels are due to the fact that growth increments become smaller with age. This fact is also illustrated in Table 56, where the differences between per cents passing in adjacent age groups are larger among the younger age groups than among the older.

The average per cents shown in Table 57 are central tendencies. Actually the items assigned to a particular age level are not all of the same difficulty. Table 55 A shows that the range was considerable. In Year VIII the range is from 55 per cent to 90 per cent. The more recent revision probably has a smaller range of difficulty at each year level, but figures are not yet available. The problem of reducing this range is a persistent one because the figures for one age group will often differ from the figures for another age group which has had somewhat different training. Nearly all the more recent revisions have used groups of children for standardization who were within one month of a particular birthday. This procedure results in greater accuracy in the location of tests at a particular age level, but does not eliminate the difficulty of securing tests which were passed by just 50 per cent of an age group.

TABLE 57

Average Per Cents of an Age Group Passing the 1916 Stanford-Binet Tests for That Age

Age	*Per Cent*
3	77.0
4	77.0
5	71.3
6	70.8
7	68.0
8	63.2
9	62.3
10	64.5
12	62.4
14	55.6
Average Adult	59.8
Superior Adult	*37.4

* Per cent of Average Adult sample.

(After Terman *et al.*, 1917, p. 158. By permission of the Editor, *Educational Psychology Monographs*.)

When the scaling of items is at last completed, the final form of the test is ready for wider use. Its correct application and scoring require a great deal of preliminary training. Persons wishing to become proficient examiners should have, in addition to a college major in psychology, at least a year of special work in observing and administering tests under careful supervision. Thorough interpretation of scores requires an even wider experience including a knowledge of social factors which may affect the results, and also a knowledge of probable sources of error. Some of the facts needed for interpretation are given below.

VII. INTERPRETATION OF MENTAL AGES

Mental age scores from Binet-type scales have the advantages of being widely referred to, fairly easily calculated, and easily understood in some respects. Since one often gets useful information from these scales, he should be aware of their main limitations. Some of the sources of error in the use of mental ages are:

A. Inequality of Steps

Nearly all studies of mental growth show that increments of growth decrease from year to year. For most of the available scales the observed differences between mental ages of 2 and 3 years are much more noticeable than the differences between mental ages of 9 and 10 years. Many of the mental age scales ignore this important fact, and their use leads to conclusions that a mental age of 8 years represents twice the ability of a mental age of 4, or that a year's growth from 1 to 2 years is equal to that from 11 to 12 years. Such statements, although common, are of very doubtful value since they require special interpretations.

B. Changes in Meaning of Mental Ages above Adult Level

It is altogether impossible to measure persons in the upper half of the adult population on ordinary mental age scales, for they surpass the average adult's score, and a mental age was originally defined as the average score made by an age group. In order to overcome this difficulty, Terman (1916) and later several others have arbitrarily extended mental age scales above the average adult level. Thus, a mental age of 20 is not the score of an average 20-year-old, but of a person who is generally in the highest 5 per cent of the 20-year-olds. A mental age in this case loses its original meaning of representing the average of a particular age group, and becomes an arbitrary point fixed above the average in such a way that the IQ's of adults will be distributed according to some hypothesis.

In assigning arbitrary MA's above the average adult, Terman and Merrill (1937) assumed that adults should have the same distribution of IQ's as the children who can be measured using true MA scores — that is, children from approximately 5 to 10 years of age. These age groups were found to have fairly normal distributions of IQ's with the mean near 100 and the standard deviation approximately 17. The scaling method for adults was as follows: Tests which were passed by less than half of the adults tested were arranged in order of difficulty as shown by the per cent passing. They were then assigned places in various adult levels, and given a certain number of months-of-growth credit so that the adult sample would have an average IQ near 100 and a standard deviation of 17. This process resulted in a Superior Adult Level I, which corresponds to a mental age of 17 years, 4 months, with six tests (see Table 53), each of which is assigned a mental age credit of 4 months. The Superior Adult Level II contains six tests, the successful completion of each of which is given credit of 5 months. The Superior Adult Level III, which corresponds to a mental age of 22 years, 10 months, has six harder tests in it, each of which is assigned credit of 6 months.

In Kuhlmann's work the extension of the (MU) point scale beyond the average adult level was made from the Heinis curve. The curve was expressed by the formula

$$y = b \left(1 - e^{-\frac{x}{d}}\right)$$

where x is age, y is the MU curve value; b is 429, the maximum value of y; e is the base of the natural logarithm; and d is a constant, 6.675. Kuhlmann computed mental unit equivalents on the Heinis curve for each year and month of mental age and located the maximum value at over 45 years. It was later found that the mean raw scores on each test had a linear relation with MU's below the adult level. The scores above the average adult level were therefore assigned equal increases in MU's corresponding to equal increases in raw scores.

These changes in the meaning of mental age above the average adult level are not as widely understood as they should be, and this misunderstanding has led to many errors in interpreting results. Moreover, the methods of assigning arbitrary mental age values are not very well agreed upon, so that various authors have devised different standards which are not interchangeable.

C. Processes Measured at Different Ages

An inspection of all the Binet-type scales shows a distinct tendency away from the use of pictures or objects and toward more abstract verbal skills with increasing MA levels. Statistical evidence that the same or similar processes are sampled at each age level is not yet available. Research is needed to show the amount of similarity or dissimilarity of processes sampled at various age levels in order to allow correct interpretation.

D. Components of Equal Mental Ages

The question has often been raised whether a mental age of 10 in a 10-year-old represents the same abilities as a mental age of 10 in a 15-year-old. The answer has been given by elaborate studies comparing the scores of retarded and normal children of the same MA's (see Table 55 B). This table shows that Items VI–1 and VI–6 are more difficult for retarded than for normal or superior children, whereas VI–2, VI–3, and VI–5 are much easier for the retarded than for the superior. Burt (1922) and Merrill (1924), summarizing the work of others as well as their own, state that there is a marked tendency for older persons in an MA group to succeed better than the younger on items which depend to some extent on muscular maturation or rote memory. On the other hand, verbal discriminations, unusual interpretations, and number relations are relatively easier for the bright than for the dull. These results indicate that there may be several fairly independent variables in the test situation, such as rote memory for familiar facts, observation and comparison activities, and number combinations.

This evidence is confirmed by many other observers. A mental age of 10 in the case of the retarded person usually represents a slow rote performance, whereas a mental age of 10 in the case of a superior boy represents success in quick observations and inferences. The speed factor, which seems important in evaluating superior mental ability, does not count very much on the usual Binet scale but it is stressed in the higher levels of Kuhlmann's Revision. Since the Binet-type tests are not designed for analytical work, they do not demonstrate very well these differences between bright and dull children. Variations in the meaning of MA's can only be overcome by designing tests in which a certain MA score will always represent a particular pattern of skill.

VIII. INTERPRETATION OF INTELLIGENCE QUOTIENTS

In order to show a person's relative rate of growth, a ratio known as the intelligence quotient, or IQ, has been widely used. This is determined by dividing the mental age by the chronological age of a person. Those who have a mental age of 12 and a chronological age of 10 would have intelligence quotients of 120. (Decimal points are usually avoided by multiplying by 100.) Intelligence quotients are interpreted either as percentages of normal development, or as indicators of the relative speed of development. Thus, a child who has an IQ of 200 has developed twice as much MA in a given time, or has developed twice as fast as the normal. Because the IQ is so widely used, its main limitations need to be examined closely. These may be considered under the following headings:

A. Similarity of Average IQ's

One criterion of comparability is seen in the average scores of various age groups. If average scores are all 100 or nearly 100, then one can be sure that persons with IQ's of 100 all stand at the middle or 50th centile of their age groups. On the 1937 revision of the Stanford-Binet Scale, the average IQ's of various age groups used for standardization actually ranged from 100 to 109. This means that an IQ of 100 represents the 50th centile in one group and approximately the 30th centile in another, since the standard deviations of these groups were all approximately 17 points (see Chapter 3). The authors found, however, that smoothed curves of average IQ's showed less variation, and it is usually the case that larger populations give more constant average scores. Hence these variations in mean IQ's are only considered to be a serious source of error in very careful studies of growth.

B. Similarity of Standard Deviations of IQ's

Frequency distributions of IQ's of various age groups have usually given curves which appear to be nearly normal in shape, with the mean at approximately 100 and the standard deviation about 17. Table 58 shows distributions of IQ's and two commonly used classifications of brightness.

There may be, however, differences in the dispersions of IQ's for various age groups. Kuhlmann (1939) reported larger SD's in preschool groups than among the older groups. Terman and Merrill (1937) attempted to keep dispersions constant at approximately 17 points by selection of items, but still there appear variations. They point out (p. 40):

> Attention, however, should be drawn to ages six and twelve, where the relatively low and high values respectively are deviations too extreme to be explained as purely chance fluctuations. The high variability at age twelve might conceivably be ascribed to the differential age of the onset of pubescence, although it has yet to be demonstrated that pubescence is significantly related to the rate of mental growth. Whether the atypical IQ variability at age six resides in the character of the sampling at that age, or whether it is perhaps an artifact of the nature of the scale at that level cannot be determined from the available data. In the lack of positive proof to the contrary, we are probably justified in assuming that the true variability is approximately constant from age to age. Repeated tests of the same subjects from early childhood to maturity will be necessary to determine whether this assumption is in accord with the facts.

TABLE 58

A. Descriptive Classification of Stanford-Binet IQ's

Range of IQ	*Descriptive Term*	*Approximate Per Cent*	*Range of Probable Adult MA*
above 150	near genius	.1	22.50 and above
130–149	very superior	3.	19.50–22.49
115–129	superior	14.	17.25–19.49
85–114	normal	66.	12.75–17.24
70–84	dull	14.	10.50–12.74
60–69	borderline	2.	9.00–10.49
40–59	moron	1.	6.00– 8.99
20–39	imbecile	—	3.00– 5.99
0–19	idiot	—	0.00– 2.99

(Bernreuter and Carr, 1938, p. 314. By permission of the Editor, *Journal of Educational Psychology.*)

B. Distributions of Composite L–M IQ's of the Standardization Group

IQ	*Classifications*	*Per Cent*
160–169	very superior	0.03
150–159		0.2
140–149		1.1
130–139	superior	3.1
120–129		8.2
110–119	high average	18.1
100–109	normal or average	23.5
90–99		23.0
80–89	low average	14.5
70–79	borderline defective	5.6
60–69	mentally defective	2.0
50–59		0.4
40–49		0.2
30–39		0.03

NOTE: The normal group in A is more inclusive than in B. Merrill pointed out that the distribution of scores in B was somewhat skewed toward the higher end of the scale.

(Merrill, 1938. By permission of the Editor, *Journal of Educational Psychology.*)

C. Calculation for Adults

When one attempts to calculate IQ's for adults, one meets a serious difficulty, because mental age gradually stops increasing during adolescence, while chronological age, of course, increases regularly. The difficulty is illustrated by the case of a person who had an MA of 16 when he was 16 years old, and hence an IQ of 100. The same person might still have an MA of 16 at the age of 24, which, if IQ's were calculated as before, would result in an IQ of 67. Such a change in IQ is undesirable and with the usual interpretation of IQ's it would be ridiculous. In order to have adults' IQ's remain constant, Terman decided to let chronological age remain constant from the time when mental age seemed to cease increasing in normal individuals.

A good deal of work has been done to find the average age of reaching maturity of intelligence or general ability, but no definite point has been found because growth ceases very gradually and it is difficult to secure good samples of the whole adult population. Moreover, scores in some skill and information tests continue to increase after others have ceased. For these reasons various authors differ in fixing the age of maximum mental growth. The results of measuring white soldiers in the United States army showed that the average recruit made approximately the same Stanford-Binet score as a youngster in school who was 13 years and 9 months of age. Terman (1916) placed mental maturity at 16 years, Pintner (1931) at 14 years, Baker (1935) at 15 years and 8 months, and Kuhlmann (1922) at 18 years, Terman and Merrill[2] at 15 years, and Kuhlmann (1939) at 16 years. These discrepancies may be overlooked in rough comparisons, but they often make it inadvisable to compare scores from one scale with those of another without some adjustment. It seems probable that there will always be serious discrepancies in estimation of the average age of reaching maturity until independent components of the skills which are to be tested are defined more accurately and measured separately (see Chapter 21).

[2] In calculating chronological age in the thirteenth, fourteenth, and fifteenth years they do not count one out of every three months and no increase is counted after the sixteenth year.

D. Constancy of IQ's

There are two reasons why great importance has been attached to constancy of the IQ of a person when tested at various ages. One is the hypothesis that an IQ represents a constant native ability to develop. None of the leaders in the field of measurement subscribe to this hypothesis without reservations, but it has often been assumed to be true by others. The constancy of obtained IQ's is mistaken for evidence that the IQ's indicate native capacity. This conclusion is justified only if environment has been held rigidly constant. In normal society, enormous differences in motivation or opportunity to develop are sometimes apparent even in the same family. Thus a constantly low IQ in a child may in some cases mean a continuously poor environment (see Chapter 24) .

The other reason for attaching great importance to the finding of fairly constant IQ's is the desire for practical predictions. In order to allow accurate predictions, an individual's IQ must be nearly the same from year to year, or vary in some regular fashion. For example, if a girl's IQ were 140 at the age of 8, 96 at the age of 10, and 115 at the age of 12, one could not predict from any test what she would rate on the next test.

Although a large number of studies have been made to determine how constant IQ's actually are in various groups of persons over various time intervals, these studies are as yet inconclusive because of the difficulties of accurate measurement which have been discussed, and because of the difficulties of studying the same children over long periods of time in a constant environment. However, a good deal of interesting work in determining the constancy of the IQ has been done which may be discussed under two headings: changes in individual IQ's, and correlations between retests. Changes in IQ's are of practical significance because they show what estimates are actually available for individual predictions. On the other hand, correlations between retests are theoretically more significant because they are not affected by certain types of scaling errors. For instance, a very high correlation between two tests may be found when the IQ's of a group have all gone up or down, or when they have a greater or smaller dispersion on the second test.

1. Changes in Size of IQ's. A number of studies have been made of changes in IQ's over periods of less than three months. When the environment is not markedly changed, most of these studies show an average improvement of about five points in IQ from the first to the second test. One of the most complete and interesting studies is that of Terman and Merrill (1937). They reported that all their correlations between IQ's on the two forms were distinctly fan-shaped, with the larger variations occurring among the higher IQ's. From combined results of ages 3 to 18 years, the average differences between IQ's on the two forms range from 2.49 for persons below 70 IQ, to 5.92 for persons above 130 IQ. The 1,291 persons whose IQ's were from 90 to 109, showed an average difference of 5.09 points. This difference means that approximately one-half of the normal group varied by more than five points and the other half by less than five points on a retest given within a few days. The mean variation of the superior groups was slightly more, and of the inferior somewhat less than five points.

Another interesting report is that of Cattell (1931), who obtained retests on the 1916 Stanford-Binet Test at intervals from 0 to 72 months. Table 59 summarizes her findings in part. The striking thing to notice is the loss of IQ in the children who were below average and the gain in those above average. This tendency is very marked in the longer intervals. No simple explanation of this was given. A similar finding for retarded children was reported by Berry (1923) and Kuhlmann (1921). Kuhlmann believed that loss in IQ among the retarded children was due to a defect in scaling, and he suggested a correction procedure which as-

TABLE 59

Median IQ Changes over Various Periods After Cattell, 1931

	Months between tests *0 to 24*		Months between tests *36 to 72*	
IQ	N	Points Changed	N	Points Changed
below 70	11	+ 4.0	15	− 7.0
70–79	41	− 0.8	38	− 3.5
80–89	75	+ 0.8	110	− 3.3
90–99	116	− 0.6	193	− 2.2
100–109	127	+ 1.1	166	+ 1.7
110–119	138	+ 1.1	101	+ 1.4
120–129	66	+ 0.8	35	+ 4.8
130–139	26	+ 1.5	13	+ 16.0
140 +	18	+ 8.5	0	—

(After Cattell, 1931, p. 547, Stanford-Binet Test. By permission of the Editor, *Journal of Educational Psychology.*)

sumes a growth curve suggested by Heinis (1926), Figure 99, page 673. The correction does seem to result in somewhat more constant IQ's for both bright and dull children but it is not widely used because it is usually small, and because Heinis' curve has not been widely accepted as typical of mental growth. Other studies of retests over long periods of time are discussed in Chapter 21, which deals with appraisals of growth.

2. CORRELATION BETWEEN RETESTS. The correlations between retests have been fairly numerous. They are well summarized by Nemzek (1933), who points out the difficulties of making direct comparisons. In general the correlations in small groups, or over longer periods, are smaller than the correlations in large groups, or over shorter periods.

Terman and Merrill (1937) reported that correlations between Forms L and M, administered within a few days, have a median of .88 for ages 2 to 6 years, and of .93 for ages above 6 years. The members of the 21 age groups were all within four weeks of a birthday or half-birthday. The spread of scores and the consequent correlations are therefore smaller than would be found in age groups with ranges larger than eight weeks. The authors also calculated correlations for IQ's from Forms L and M for five IQ groups in single age groups (p. 46):

130 and over	.898
111–129	.912
90–109	.924
70–89	.945
Below 70	.982

The lower IQ's appear to be more stable in score than the higher, but all are highly predictable.

IX. SUBSTITUTES FOR MENTAL AGES AND INTELLIGENCE QUOTIENTS

These considerations of the interpretation of scores lead to the conclusion that a substitute or supplement for mental age and IQ scores would be desirable if it could avoid difficulties in scaling and interpretation. Two indices have come into common use, *centiles* and *standard scores.*

The centile gives a clear picture of where a person stands in his own age group. Thus, an adult with a centile score of 90 would surpass the lowest 89 per cent of adults and be surpassed by the highest 10 per cent of his group. Likewise, a 4-year-old whose centile score is 90, would surpass the lowest 89 per cent of his group, and the highest 10 per cent would surpass him. One objection to centile scores is that they do not represent equal steps in a scale, see Chapter 3. Thus, the difference between scores of the 5th and 10th centile is much more obvious to an observer than that between the 45th and 50th centiles. The explanation is that the usual frequency curve shows a bunching of scores near the middle. This inequality of centiles, however, leads to only slight errors, and centiles are coming into common use for all age groups.

Another substitute for MA's and IQ's which has the advantage of giving units of measurement which are equal in appearance is the standard score, described in Chapter 3. This is not quite so easy to learn to interpret as the centile, but it seems just as easy to interpret as an IQ and has a wider range of application. Thus, standard scores are commonly used for many measures which involve motor as well as mental skills, and for all sorts of achievement tests (see Table 10 and Figure 16). Standard scores seem to be already preferred to IQ's when dealing with adults, as shown by the frequency of their use among college and employment bureau records.

Both standard and centile scores have the advantage of comparing a person's score with those of the same age group. Centiles and standard scores cannot, however, be substituted for MA's and IQ's indiscriminately. The various indices are indications of different relationships, and each should be used when it seems most appropriate. Each has limitations in meaning which are supplemented by the others. Table 10 shows how all four scores are related to each other. Such tables make it possible to secure the desired index for a test score very quickly.

The difficulties which surround the use of MA's and IQ's led to the appointment of a committee by the Clinical Section of the American Psychological Association to study the matter. This committee's report, which was approved in September, 1937, states in part:

> 1. In cases where reports of test scores continue to be made in terms of age levels ("mental age," "educational age," "reading age," etc.) or of quotients derived from these ("intelligence quotient," "achievement quotient," and the like) the name of the test used and the ages of the subjects tested should always be given.
>
> 2. The authors of standard tests already in use should be requested by the Section to publish tables which make it possible to translate test data into scores which show the degree of deviation from the published norms, and publishers should not accept any tests for publication unless the authors furnish such tables.

3. In making reports to parents and others without psychological training, mental ages, intelligence quotients, etc., should seldom be used, and should give way to the use of scores which give some measure of deviation.

X. PRINCIPAL USES

Binet-type tests have been extensively used in many types of investigation. Their significance in appraisals of innate factors is discussed in Chapter 24, and of individual growth in Chapter 21. Binet-type tests have also had wide usage in clinical and vocational appraisals of abilities, which are described in various publications.

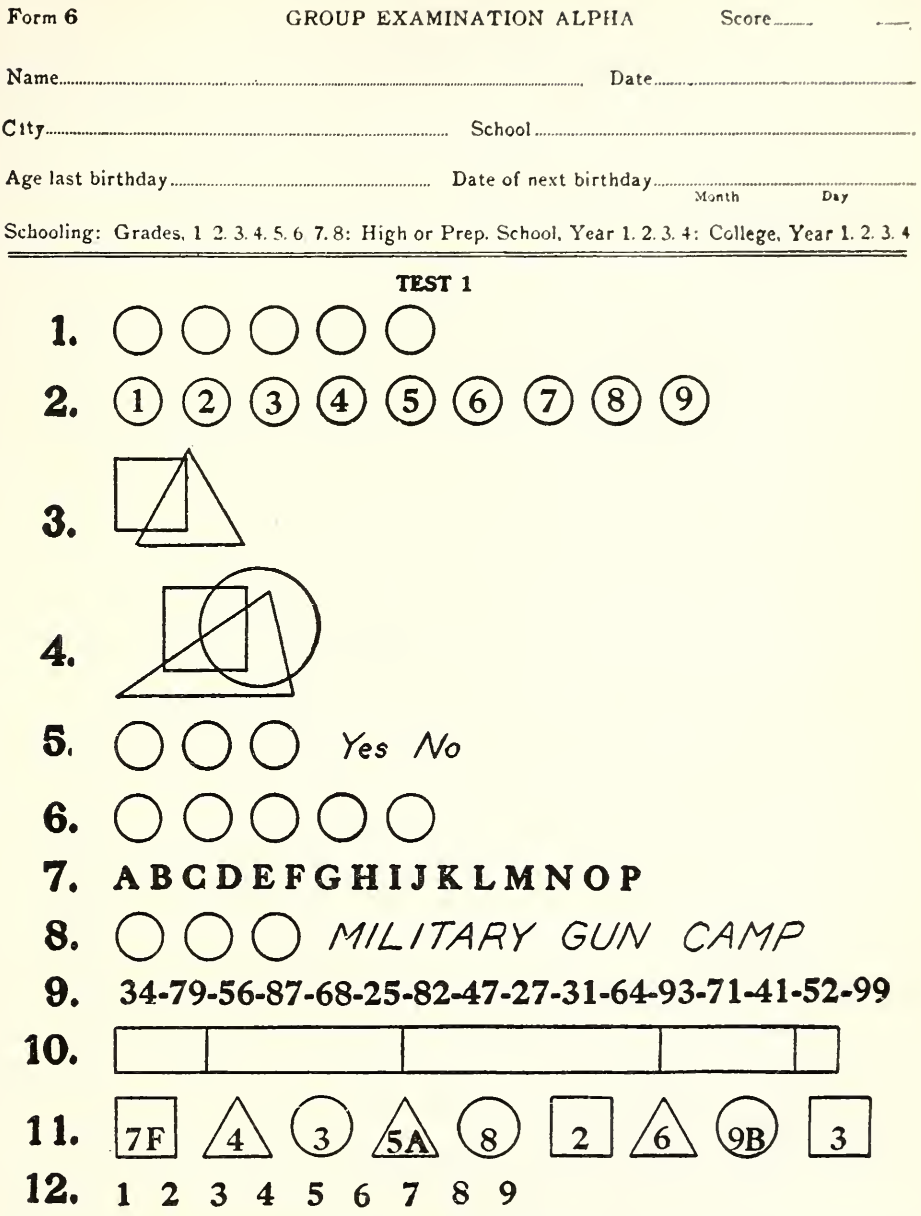
11

GROUP INTELLIGENCE TESTS

I. INTRODUCTION

During the same period that individual tests of the Binet type were being developed there was also an active growth in the design of mental tests for use with groups. All types of tests which are now widely used seem to have been fairly well developed in form before 1910. For instance, careful methods for measuring memory span were described by Jacobs (1887) and refined by Ebert and Neumann (1905). In 1889 Cattell and Bryant tried out a number of tests of both controlled and uncontrolled association, which were later developed by Jastrow (1891). The use of standard arithmetic computation tests in the study of association processes was begun in 1895 by Oehrn and also by Kraeplin. During 1897 Ebbinghaus published an enthusiastic account of a sentence completion test as a real test of intelligence. Kirkpatrick (1891) described a rather difficult vocabulary test.

In 1903 Swift published his work on interpretation of fables or proverbs. Whipple (1908) published fairly elaborate vocabulary and anagram tests and in 1909 a range of information test. In 1910 an Englishman named Brown



Form 6 GROUP EXAMINATION ALPHA Score

Name Date

City School

Age last birthday Date of next birthday Month Day

Schooling: Grades, 1 2. 3. 4. 5. 6 7. 8: High or Prep. School, Year 1. 2. 3. 4: College, Year 1. 2. 3. 4

TEST 1

1. ○ ○ ○ ○ ○

2. ① ② ③ ④ ⑤ ⑥ ⑦ ⑧ ⑨

3.

4.

5. ○ ○ ○ *Yes No*

6. ○ ○ ○ ○ ○

7. **A B C D E F G H I J K L M N O P**

8. ○ ○ ○ *MILITARY GUN CAMP*

9. **34-79-56-87-68-25-82-47-27-31-64-93-71-41-52-99**

10.

11. 7F 4 3 5A 8 2 6 9B 3

12. **1 2 3 4 5 6 7 8 9**

(This test and those following on pages 300–306 reproduced by permission of the C. H. Stoelting Co., and Henry Holt and Co.)

wrote a book on the essentials of mental measurement, and Cyril Burt (1909) reported elaborate experimental tests of higher mental processes. In the United States, Woodworth and Wells (1911) reported their famous association tests

TEST 2

Get the answers to these examples as quickly as you can.
Use the side of this page to figure on if you need to

SAMPLES 1 How many are 5 men and 10 men?........Answer (**15**)
SAMPLES 2 If you walk 4 miles an hour for 3 hours, how far do you walk?........................Answer (**12**)

1 How many are 40 guns and 6 guns?..................Answer ( )
2 If you save $6 a month for 5 months, how much will you save?...Answer ( )
3 If 32 men are divided into squads of 8, how many squads will there be?..Answer ( )
4 Mike had 11 cigars. He bought 3 more and then smoked 6. How many cigars did he have left?..................Answer ( )
5 A company advanced 6 miles and retreated 3 miles. How far was it then from its first position?....................Answer ( )
6 How many hours will it take a truck to go 48 miles at the rate of 4 miles an hour?...................................Answer ( )
7 How many pencils can you buy for 40 cents at the rate of 2 for 5 cents?...Answer ( )
8 A regiment marched 40 miles in five days. The first day they marched 9 miles, the second day 6 miles, the third 10 miles, the fourth 9 miles. How many miles did they march the last day?...Answer ( )
9 If you buy 2 packages of tobacco at 8 cents each and a pipe for 55 cents, how much change should you get from a two-dollar bill?...Answer ( )
10 If it takes 8 men 2 days to dig a 160-foot drain, how many men are needed to dig it in half a day?......................Answer ( )
11 A dealer bought some mules for $900. He sold them for $1,000, making $25 on each mule. How many mules were there?.Answer ( )
12 A rectangular bin holds 600 cubic feet of lime. If the bin is 10 feet wide and 5 feet deep, how long is it?..............Answer ( )
13 A recruit spent one-eighth of his spare change for post cards and four times as much for a box of letter paper, and then had 60 cents left. How much money did he have at first?..Answer ( )
14 If 2½ tons of hay cost $20, what will 4½ tons cost?....Answer ( )
15 A ship has provisions to last her crew of 600 men 6 months. How long would it last 800 men?...................Answer ( )
16 If a train goes 200 yards in 10 seconds, how many feet does it go in a fifth of a second?.............................Answer ( )
17 A U-boat makes 10 miles an hour under water and 20 miles on the surface. How long will it take to cross a 100-mile channel, if it has to go three-fifths of the way under water?.....Answer ( )
18 If 214 squads of men are to dig 4,066 yards of trench, how many yards must be dug by each squad?............Answer ( )
19 A certain division contains 2,000 artillery, 15,000 infantry, and 1,000 cavalry. If each branch is expanded proportionately until there are in all 19,800 men, how many will be added to the artillery? ....Answer ( )
20 A commission house which had already supplied 1,897 barrels of apples to a cantonment delivered the remainder of its stock to 28 mess halls. Of this remainder each mess hall received 47 barrels. What was the total number of barrels supplied?.Answer ( )

which included verbal analogies, opposites, part-whole, agent-action, species-genus, and hard directions. In 1913 Pyle reported an elaborate set of examinations of school children that resulted in age norms.

In 1914 Whipple published a two-volume manual of mental and physical tests which included fifty-one tests with

TEST 3

This is a test of common sense. Below are sixteen questions. Three answers are given to each question. You are to look at the answers carefully; then make a cross in the square before the best answer to each question, as in the sample:

SAMPLE
Why do we use stoves? Because
☐ they look well
☒ they keep us warm
☐ they are black

Here the second answer is the best one and is marked with a cross. Begin with No. 1 and keep on until time is called.

1 If plants are dying for lack of rain, you should
☐ water them
☐ ask a florist's advice
☐ put fertilizer around them

2 A house is better than a tent, because
☐ it costs more
☐ it is more comfortable
☐ it is made of wood

3 Why does it pay to get a good education? Because
☐ it makes a man more useful and happy
☐ it makes work for teachers
☐ it makes demand for buildings for schools and colleges

4 If the grocer should give you too much money in making change, what is the right thing to do?
☐ buy some candy of him with it
☐ give it to the first poor man you meet
☐ tell him of his mistake

5 Why should food be chewed before swallowing?
☐ it is better for the health
☐ it is bad manners to swallow without chewing
☐ chewing keeps the teeth in condition

6 If you saw a train approaching a broken track you should
☐ telephone for an ambulance
☐ signal the engineer to stop the train
☐ look for a piece of rail to fit in

7 If you are lost in a forest in the daytime, what is the thing to do?
☐ hurry to the nearest house you know of
☐ look for something to eat
☐ use the sun or a compass for a guide

8 It is better to fight than to run, because
☐ cowards are shot
☐ it is more honorable
☐ if you run you may get shot in the back

☞ Go to No. 9 above

9 Why are warships painted gray? Because gray paint
☐ is cheaper than other colors
☐ is more durable than other colors
☐ makes the ships harder to see

10 Why should all parents be made to send their children to school? Because
☐ it prepares them for adult life
☐ it keeps them out of mischief
☐ they are too young to work

11 The reason that many birds sing in the spring is
☐ to let us know spring is here
☐ to attract their mates
☐ to exercise their voices

12 Gold is more suitable than iron for making money because
☐ gold is pretty
☐ iron rusts easily
☐ gold is scarcer and more valuable

13 The cause of echoes is
☐ the reflection of sound waves
☐ the presence of electricity in the air
☐ the presence of moisture in the air

14 We see no stars at noon because
☐ they have moved around to the other side of the earth
☐ they are so much fainter than the sun
☐ they are hidden behind the sky

15 Some men lose their breath on high mountains because
☐ the wind blows their breath away
☐ the air is too rare
☐ it is always cold there

16 Why do some men who could afford to own a house live in a rented one? Because
☐ they don't have to pay taxes
☐ they don't have to buy a rented house
☐ they can make more by investing the money the house would cost

directions and norms, and about five hundred references to technical reports. Since that time the production of technical reports and test revisions has been voluminous, and many refinements have been made in administration, scoring, and scaling procedures.

Most of the early examiners, like those of today, followed

TEST 4

If the two words of a pair mean the same or nearly the same, draw a line under *same*. If they mean the opposite or nearly the opposite, draw a line under *opposite*. If you cannot be sure guess. The two samples are already marked as they should be.

SAMPLES good—bad . . . . same—opposite
SAMPLES little—small . . . . same—opposite

1 cold—hot . . . . same—opposite 1
2 long—short . . . . same—opposite 2
3 bare—naked . . . . same—opposite 3
4 joy—happiness . . . . same—opposite 4
5 find—lose . . . . same—opposite 5

6 shrill—sharp . . . . same—opposite 6
7 minus—plus . . . . same—opposite 7
8 grim—stern . . . . same—opposite 8
9 careless—anxious . . . . same—opposite 9
10 crude—coarse . . . . same—opposite 10

11 commend—approve . . . . same—opposite 11
12 linger—loiter . . . . same—opposite 12
13 agony—bliss . . . . same—opposite 13
14 defective—normal . . . . same—opposite 14
15 competent—qualified . . . . same—opposite 15

16 knave—villain . . . . same—opposite 16
17 null—void . . . . same—opposite 17
18 wax—wane . . . . same—opposite 18
19 adversary—colleague . . . . same—opposite 19
20 altruistic—egotistic . . . . same—opposite 20

21 furtive—sly . . . . same—opposite 21
22 any—none . . . . same—opposite 22
23 asunder—apart . . . . same—opposite 23
24 deplete—exhaust . . . . same—opposite 24
25 superfluous—essential . . . . same—opposite 25

26 recoup—recover . . . . same—opposite 26
27 celibate—married . . . . same—opposite 27
28 recant—disavow . . . . same—opposite 28
29 avarice—cupidity . . . . same—opposite 29
30 aggrandize—belittle . . . . same—opposite 30

31 decadence—decline . . . . same—opposite 31
32 nullify—annul . . . . same—opposite 32
33 ambiguous—equivocal . . . . same—opposite 33
34 agglomerate—scatter . . . . same—opposite 34
35 plenary—complete . . . . same—opposite 35

36 suavity—asperity . . . . same—opposite 36
37 perfunctory—meticulous . . . . . . same—opposite 37
38 lugubrious—maudlin . . . . same—opposite 38
39 desuetude—disuse . . . . same—opposite 39
40 adventitious—accidental . . . . . . same—opposite 40

the hypothesis that persons were possessed of a general faculty called intelligence, which could be measured by a variety of mental tests. They usually wished to appraise intelligence for practical predictions of some sort. They selected only those items for an intelligence test which showed fairly high correlations with some criterion of intelligence and low

TEST 5

The words A EATS COW GRASS in that order are mixed up and don't make a sentence; but they would make a sentence if put in the right order: A COW EATS GRASS, and this statement is true.

Again, the words HORSES FEATHERS HAVE ALL would make a sentence if put in the order ALL HORSES HAVE FEATHERS, but this statement is false.

Below are twenty-four mixed-up sentences. Some of them are true and some are false. When I say "go," take these sentences one at a time. Think what each would say if the words were straightened out, but don't write them yourself. Then, if what it would say is true, draw a line under the word "true"; if what it would say is false, draw a line under the word "false." If you can not be sure, guess. The two samples are already marked as they should be. Begin with No. 1 and work right down the page until time is called.

SAMPLES
a eats cow grass....................true..false
horses feathers have all..............true..false

1 cows milk give.......................true..false 1
2 write are with to pencils.....................true..false 2
3 are and apples long thin.......................true..false 3
4 east the in rises sun the......................true..false 4
5 months warmest are summer the..............true..false 5
6 wood made carpets are of always..............true..false 6
7 known elephant animal an is smallest the.......true..false 7
8 water cork on float will not...................true..false 8
9 vote children 21 cannot under.................true..false 9
10 battleships on seldom sails used are...........true..false 10
11 four hundred all pages contain books...........true..false 11
12 iron paper made of is filings..................true..false 12
13 pays cautious it be to often..................true..false 13
14 a general not major a and rank same the of are..true..false 14
15 Washington canal 1776 Panama the in built....true..false 15
16 never deeds rewarded be should good..........true..false 16
17 will live bird no forever.....................true..false 17
18 gases the in Mohawks fighting used poisonous...true..false 18
19 friends in us disaster often false desert.........true..false 19
20 external deceptive never appearances are.......true..false 20
21 size now of guns use are great in.............true..false 21
22 happiness lists great casualty cause............true..false 22
23 always sleeplessness clear causes a conscience....true..false 23
24 inflict men pain needless cruel sometimes.......true..false 24

correlations with each other. An important application of this method of selection was the development of United States Army mental tests in 1917.

Approximately one and three-quarters million soldiers were tested during 1917 and 1918 by one or both of the forms developed at that time. The tests have since been

TEST 6

SAMPLES	2	4	6	8	10	12	*14*	*16*
	9	8	7	6	5	4	*3*	*2*
	2	2	3	3	4	4	*5*	*5*
	1	7	2	7	3	7	*4*	*7*

Look at each row of numbers below, and on the two dotted lines write the two numbers that should come next.

2	3	4	5	6	7	.....	.....
5	10	15	20	25	30	.....	.....
10	9	8	7	6	5	.....	.....
6	9	12	15	18	21	.....	.....
8	8	6	6	4	4	.....	.....
3	7	11	15	19	23	.....	.....
9	1	7	1	5	1	.....	.....
25	25	21	21	17	17	.....	.....
4	5	8	9	12	13	.....	.....
21	18	16	13	11	8	.....	.....
1	2	4	8	16	32	.....	.....
3	4	6	9	13	18	.....	.....
12	14	13	15	14	16	.....	.....
25	24	22	21	19	18	.....	.....
16	12	15	11	14	10	.....	.....
16	8	4	2	1	½	.....	.....
15	16	14	17	13	18	.....	.....
1	4	9	16	25	36	.....	.....
21	18	16	15	12	10	.....	.....
4	8	10	20	22	44	.....	.....

widely applied to industrial, prison, and school groups. They have also been widely copied in both form and idea. They will be briefly described.

TEST 7

SAMPLES
sky—blue :: grass— **table green warm big**
fish—swims :: man— **paper time walks girl**
day—night :: white— **red black clear pure**

In each of the lines below, the first two words are related to each other in some way. What you are to do in each line is to see what the relation is between the first two words, and underline the word in heavy type that is related in the same way to the third word. Begin with No. 1 and mark as many sets as you can before time is called.

---

1 dog—bark :: cat— **chair mew fire house** 1
2 foot—man :: hoof— **corn tree cow hoe** 2
3 dog—puppy :: cat— **kitten dog tiger horse** 3
4 wash—face :: sweep— **clean broom floor straw** 4
5 door—house :: gate— **swing hinges yard latch** 5

6 water—fish :: air— **spark man blame breathe** 6
7 white—black :: good— **time clothes mother bad** 7
8 boy—man :: lamb— **sheep dog shepherd wool** 8
9 roof—house :: hat— **button shoe straw head** 9
10 camp—safe :: battle— **win dangerous field fight** 10

11 straw—hat :: leather— **shoe bark coat soft** 11
12 pan—tin :: table— **chair wood legs dishes** 12
13 left—right :: west— **south direction east north** 13
14 floor—ceiling :: ground— **earth sky hill grass** 14
15 cold—ice :: heat— **wet cold steam stars** 15

16 hat—head :: thimble— **sew cloth finger hand** 16
17 Monday—Tuesday :: Friday— **week Thursday day Saturday** 17
18 lead—bullet :: gold— **paper coin silver copper** 18
19 skin—body :: bark— **tree dog bite leaf** 19
20 cannon—large :: rifle— **ball small bore shoot** 20

21 cellar—attic :: bottom— **well tub top house** 21
22 man—arm :: tree— **shrub limb flower bark** 22
23 suitcase—clothing :: purse— **purchase money string stolen** 23
24 knitting—girls :: carpentry— **trade houses boys lumber** 24
25 arteries—body :: railroads— **country train crossing accident** 25

26 ocean—pond :: deep— **sea well shallow steep** 26
27 revolver—man :: sting— **gun hurt bee hand** 27
28 engineer—chauffeur :: locomotive— **iron stack engine auto** 28
29 terrier—dog :: Jersey— **City cow horse State** 29
30 airplane—air :: submarine— **dive engine ship water** 30

31 esteem—friends :: despise— **forsake detest enemies people** 31
32 hospital—patient :: prison— **cell criminal bar jail** 32
33 tears—laughter :: sorrow— **joy distress funeral sad** 33
34 yes—no :: affirmative— **win debate deny negative** 34
35 establish—abolish :: begin— **work year end commence** 35

36 order—confusion :: peace— **part treaty war enemy** 36
37 education—ignorance :: wealth— **poverty riches health comfort** .. 37
38 10—100 :: 1000— **money 10000 20000 wealth** 38
39 imitate—copy :: invent— **study Edison machine originate** 39
40 historian—facts :: novelist— **fiction Dickens writer book** 40

II. THE UNITED STATES ARMY TESTS

As criteria for defining intelligence, the psychologists in charge of the work of testing the soldiers decided to use combined scores of (*a*) formal school accomplishment, (*b*) scores on the Stanford-Binet test, and (*c*) ratings by

TEST 8

Notice the sample sentence:

People **hear** with the eyes **ears** nose **mouth**

The correct word is **ears**, because it makes the truest sentence.

In each of the sentences below you have four choices for the last word. Only one of them is correct. In each sentence draw a line under the one of these four words which makes the truest sentence. If you can not be sure, guess. The two samples are already marked as they should be.

SAMPLES { People hear with the eyes ears nose mouth
France is in Europe Asia Africa Australia

1 Boston is in Connecticut Rhode Island Maine Massachusetts.................... 1
2 Euchre is played with dice rackets cards pins.................................. 2
3 The Arabian is a kind of horse goat cow sheep.................................. 3
4 The most prominent industry of Milwaukee is fish brewing flour automobiles...... 4
5 Turquoise is usually yellow red green blue...................................... 5
6 The Leghorn is a kind of cow horse fowl granite................................. 6
7 Arthur Brisbane is famous as a newspaper man comic artist athlete actor.......... 7
8 Shoes are made by Swift & Co. Smith & Wesson W. L. Douglas Babbitt Co...... 8
9 Blanche Sweet is known as a writer singer suffragist movie actress................ 9
10 "The makings of a nation" is an advertisement of a tobacco flour beer health food.. 10
11 Country Gentleman is a kind of wheat corn hay oats............................ 11
12 The artichoke is a vegetable fish lizard snake................................. 12
13 Yale University is at New Haven Annapolis Ithaca Cambridge.................... 13
14 Tokio is a city of India China Egypt Japan..................................... 14
15 Diamonds are obtained from mines reefs elephants oysters...................... 15
16 Rodin is famous as a poet painter sculptor composer........................... 16
17 The chameleon is a bird reptile insect fish.................................... 17
18 The thyroid is in the shoulder neck head abdomen............................. 18
19 Dioxygen is a disinfectant food product patent medicine tooth paste.............. 19
20 The U. S. S. Michigan is a destroyer monitor submarine battleship............... 20
21 The cutlass is a kind of sword musket cannon pistol............................ 21
22 The Corona is a kind of phonograph multigraph adding machine typewriter........ 22
23 Indigo is a food drink color fabric.. 23
24 The xylophone is used in lithography music stenography book-binding............ 24
25 Madras is a drink fabric food dance.. 25
26 The author of "The Scarlet Letter" is Hawthorne Poe Stevenson Kipling.......... 26
27 John Wesley was most famous in literature science war religion................ 27
28 The Delco System is used in plumbing filing ignition cataloguing................ 28
29 Rubber is obtained from ore petroleum trees hides.............................. 29
30 Darwin was most famous in literature science war politics...................... 30
31 Falstaff appears in Romola Vanity Fair Oliver Twist Henry IV.................. 31
32 The number of a Korean's legs is two four six eight............................. 32
33 Perjury is a term used in pedagogy law theology medicine........................ 33
34 A tedder is used in farming fishing hunting athletics............................ 34
35 Slice is a term used in bowling golf tennis football............................. 35
36 The Battle of Lexington was fought in 1620 1775 1812 1864..................... 36
37 The kilowatt is used in measuring rainfall wind power electricity water power........ 37
38 The Buick car is made in Toledo Flint Buffalo Detroit........................... 38
39 Among the allies of Germany is Bulgaria Norway Rumania Portugal................ 39
40 An eight-sided figure is called a trapezium scholium parallelogram octagon......... 40

officers. Groups of soldiers for whom these criteria were available were given batteries of 13 preliminary verbal tests. Four of the preliminary tests were taken from the work of Otis, who generously placed them at the disposal of the committee. Several tests show great similarities to those in the Woodworth-Wells series. Binet's and Thurstone's tests were also used or adapted in making up these Army tests.

The correlations of total preliminary test scores with officers' ratings of intelligence ranged from approximately .50 to .70; with Stanford-Binet Test score, .80 to .90; with Trabue Language Completion Scales, .72; with schooling, .75; and with the Beta Test, .80. The lowest correlations of the separate subtests with total weighted scores (Yerkes, 1921, p. 541) were approximately .65 for tests of *oral directions, memory span, disarranged sentences,* and *practical judgment.* The highest correlations were approximately .85 for tests of *arithmetic problems, verbal opposites, information, verbal analogies,* and *number comparisons.* The mean of the correlations between subtests for a sample of 895 soldiers was approximately .61, and the subtests which correlated most highly with the total scores also showed the highest correlations with the other subtests. These results led empirically to the conclusion, which is also mathematically obtainable, that it is impossible to secure subtests which will correlate highly with a criterion, and nearly zero with each other. Subtests which had degrees of correlation with each other and with total scores similar to those shown above were finally selected and combined into a test called the United States Army Alpha Test. In constructing the test forms, the makers also considered a number of practical factors. The test should:

1. Be adapted for use with large groups of persons who had wide differences in ability (The final test items ranged from those which were answered by nearly 99 per cent of a group of adults to those answered correctly by about 1 per cent.)
2. Have a number of equivalent forms to prevent cheating, or coaching, or marked practice effects (Five forms of the Alpha tests which were found to give very similar results were prepared.)
3. Be arranged for ease and accuracy of scoring by clerical workers (This called for a minimum of writing. The answers

were usually a single number or a check mark indicating a particular choice.)

4. Render clues of malingering during an examination (The tests did not succeed in giving accurate answers to this problem.)

5. Be interesting

6. Be short (The total working time of the Alpha test was limited to 24 minutes.)

A. The Alpha Test

This is a paper and pencil battery with eight subtests, each placed on a separate page of a booklet and allotted a special time limit. The first is a test of span of auditory attention. The examiner reads directions which the soldiers follow by making lines or numbers on the prepared items shown on page 299. The directions become more complicated toward the end of the test. Thus, the directions for Item 2 are:

> Attention! Look at 2 where the circles have numbers in them. When I say "Go" draw a line from circle 1 to circle 4 that will pass *above* circle 2 and *below* circle 3. – GO! (Allow not over 5 seconds.)

and for Item 11:

> Attention! Look at 11. When I say "Go" draw a line through every even number that is not in a square, and also through every odd number that is in a square with a letter. – GO! (Allow not over 25 seconds.)

The second test is a five-minute test of 20 arithmetic problems. The third test involves common sense or practical judgment, one-and-a-half minutes being allowed for the 16 items. The fourth test allows one and one-half minutes to check 40 pairs of words to show whether they are the same or opposite. The fifth test allows two minutes to rearrange 24 sentences that had been disarranged in a random fashion. The sixth test requires the completion of 20 number series, allowing three minutes. The seventh test consists of 40 verbal analogies, with a working time of three minutes. The eighth test allows four minutes to check 40 multiple-choice items of miscellaneous information. These tests are printed on pages 299 to 306.

A short interval of time is allowed between tests to consider samples of the next test. The scores are the number of correct items on all except pages 302 and 303 with which the right-minus-wrong score is used. The total possible score is 212.

B. The Beta Test

In order to provide a test for men who could not read English, a selection and standardizing procedure similar to that of the Alpha Test was followed, beginning with fifteen nonverbal tests. The result is called the United States Army Beta Test.

This is also a paper and pencil test. It was designed so that it could be demonstrated largely by pantomime and without many words. Before each test the examiner and a demonstrator showed on a large blackboard how the work was to be done. The tests are printed on pages 311 to 317.

The first test consists of 5 mazes, of which the first two were traced on the blackboard by the demonstrator. Then, when the soldiers understood what was wanted, they were told, "All right, go ahead. Do it. Hurry up." The idea of working fast was impressed on those who were working slowly because only two minutes were allowed. The second test requires that 16 pictures of piles of cubes be viewed, and the number of cubes in each be written down, two and one-half minutes. The third test is a nonverbal series completion in which the pattern of *x's* and *o's* is to be completed in each line according to the way it is printed at the start of the line with a time allowance of one and three-quarters minutes. In the fourth test the examinee is required to associate symbols with numbers on a sheet according to a code placed at the top of the page for two minutes. The fifth test allows three minutes for comparing pairs of numbers and marking with an *x* those pairs that are different. The sixth test allows three minutes for drawing in missing parts of 20 printed pictures. In the seventh test 10 rather easy paper-form-board problems are allowed two and one-half minutes. The task is to draw lines to show how the small pieces would fit into the large figure.

The score is the number of correct items except for page 314 with which one-third the number right is used, owing to the large number of items; and page 315, with which the score is right-minus-wrong.

Most of the subtests of the Alpha and Beta were included in an elaborate factorial analysis by Thurstone discussed below; so their dominant processes will not be described now. The authors of the tests generally pointed out that reasoning, perceiving, and learning are important, and hence the tests are probably good indicators of intelligence.

Through the tabulation of data from large samples of various groups of soldiers, norms and mental age equivalents for scores, as shown in Table 60, were established.

Table 60 shows that the mean mental age of a group of 653 white soldiers was approximately 13 years, 9 months. Since Terman (1916) had placed the mean mental age at 16 years, the difference between these two means of two years caused considerable comment. However, the Army

TABLE 60

Norms of United States Army Tests

Letter	*Alpha*		*Beta*	*Stanford-Binet*
Grade	*Score*	*Per Cent*	*Score*	*MA*
A	135–212	4.1	100–118	18.0–19.5
B	105–134	8.0	90–99	16.5–17.9
C +	75–104	15.0	80–89	15.0–16.4
C	45–74	25.0	65–79	13.0–14.9
C –	25–44	34.8	45–64	11.0–12.9
D	15–24	17.1	20–44	9.5–10.9
E	0–14	7.0	0.19	0–9.4
N	93,973	93,973	2,290	653

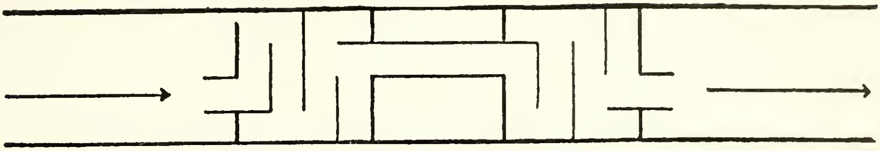
(Arranged from Yoakum, Yerkes, 1920, p. 134. By permission of Henry Holt and Co.)

Form O

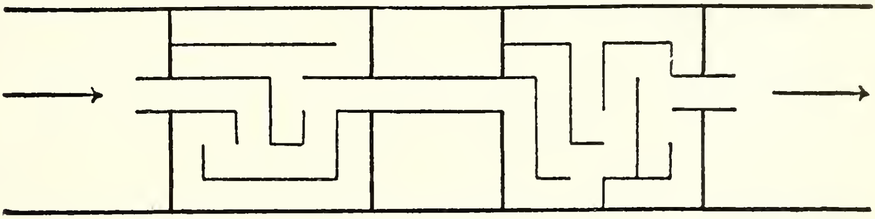
GROUP EXAMINATION BETA

Test 1

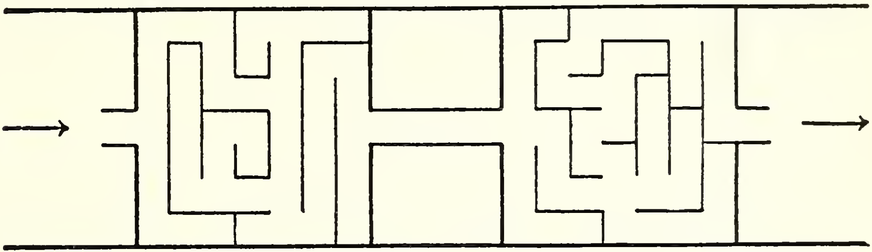
1.



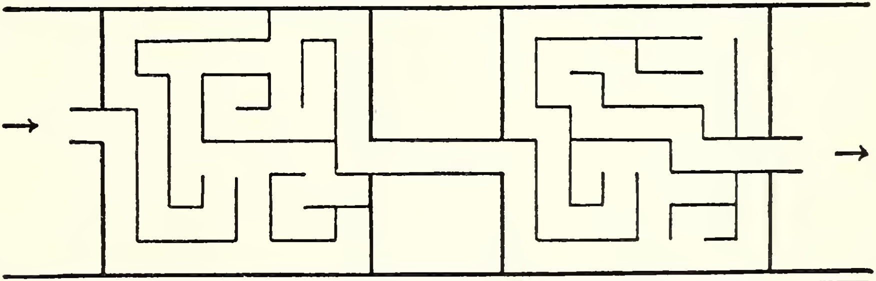
2.



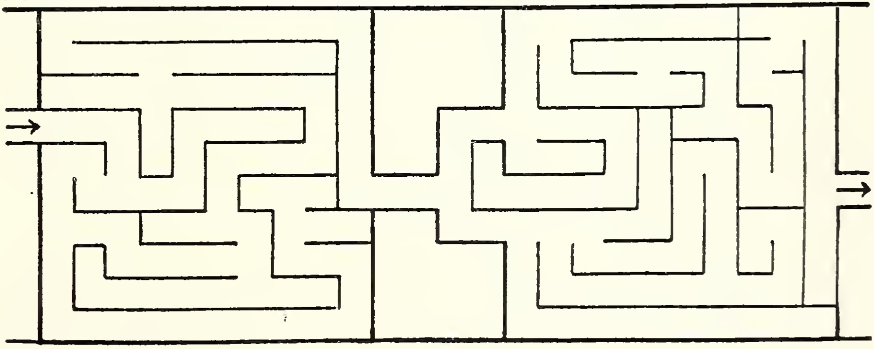
3.



4.

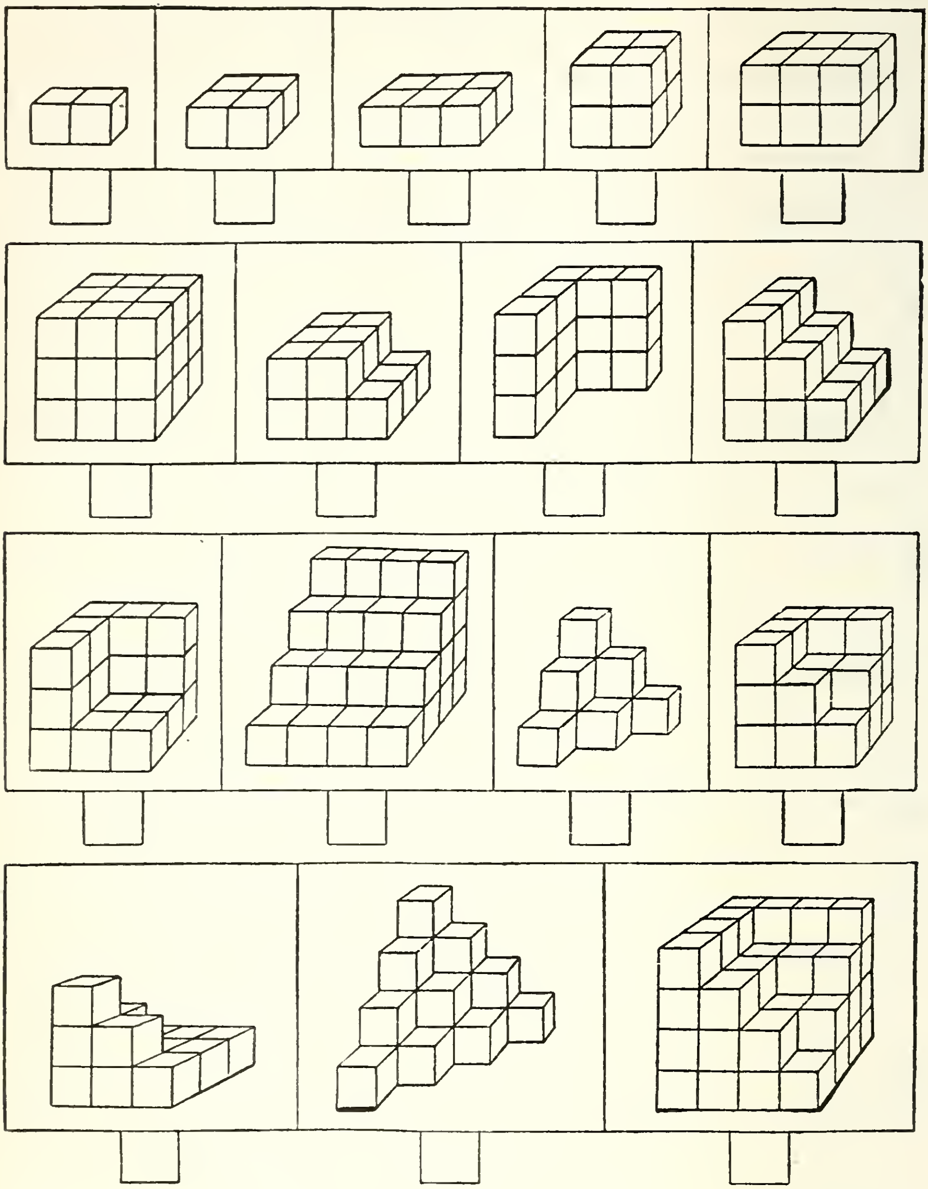


5.



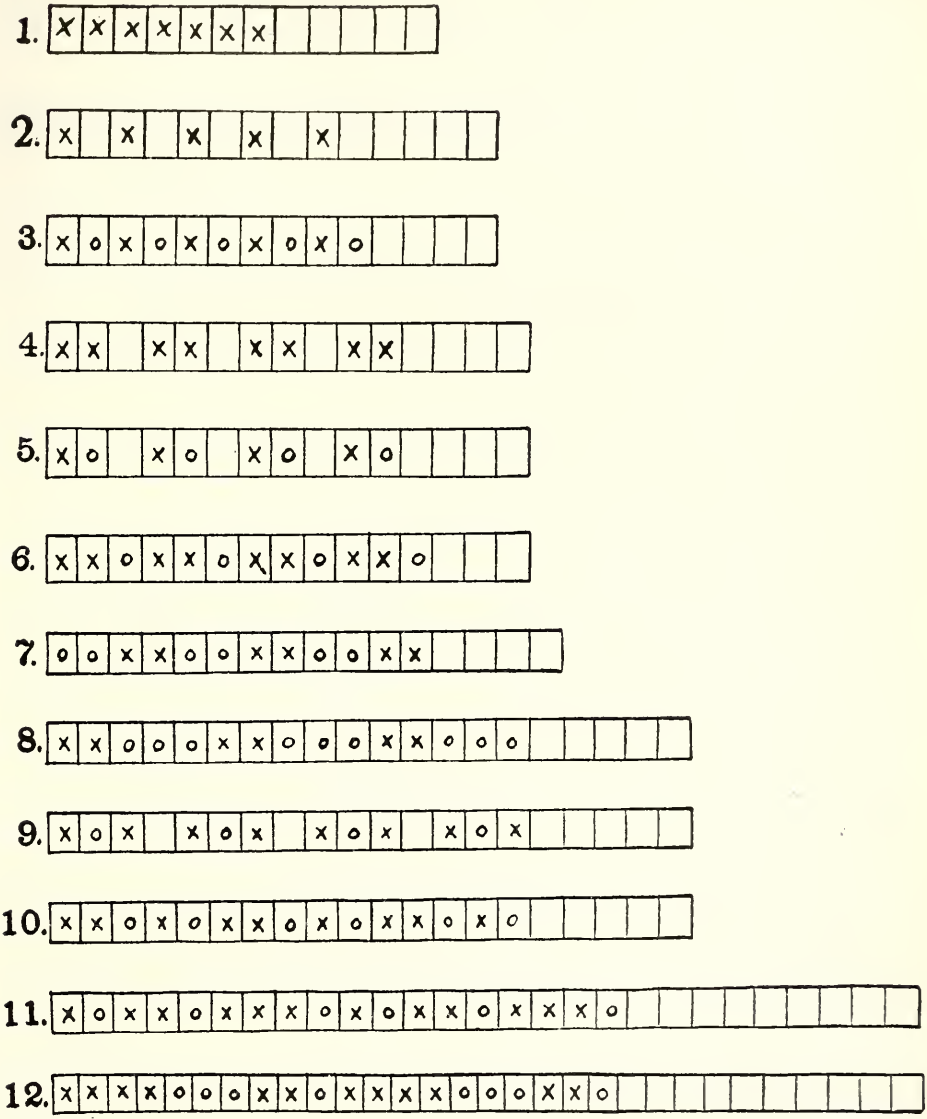
(This test and those following on pages 312–317 reproduced by permission of the C. H. Stoelting Co., and Henry Holt and Co.)

Test 2



sample was drawn from all walks of life, whereas Terman's group contained a larger proportion of high school students and business men. There is also some evidence that persons who have been out of school for four or five years respond to such mental tests less effectively than they did while in school. Terman, together with many other authorities, has

Test 3



now fixed the age of mental maturity at approximately 15 years.

During 1918 the Alpha Test Scores were found to be serviceable in discriminating the satisfactory from the unsatisfactory soldiers. Finally, Alpha Scores were used to secure similar distributions of mental ability among com-

Test 4

1	2	3	4	5	6	7	8	9
−	И	⊐	L	U	O	∧	X	=

1.

3	1	2	1	3	2	1	4	2	3	5	2	9	1	4

2.

6	3	1	5	4	2	7	6	3	8	7	2	9	5	4

3.

6	3	7	2	8	1	9	5	8	4	7	3	6	9	5

4.

1	9	2	8	3	7	4	6	5	9	4	8	5	7	6

5.

9	3	8	6	4	1	5	7	2	6	2	4	8	1	3

6

4	9	5	1	7	5	2	6	9	3	7	8	4	1	8

panies and to eliminate the unfit, for this practice was found to accelerate the training program.

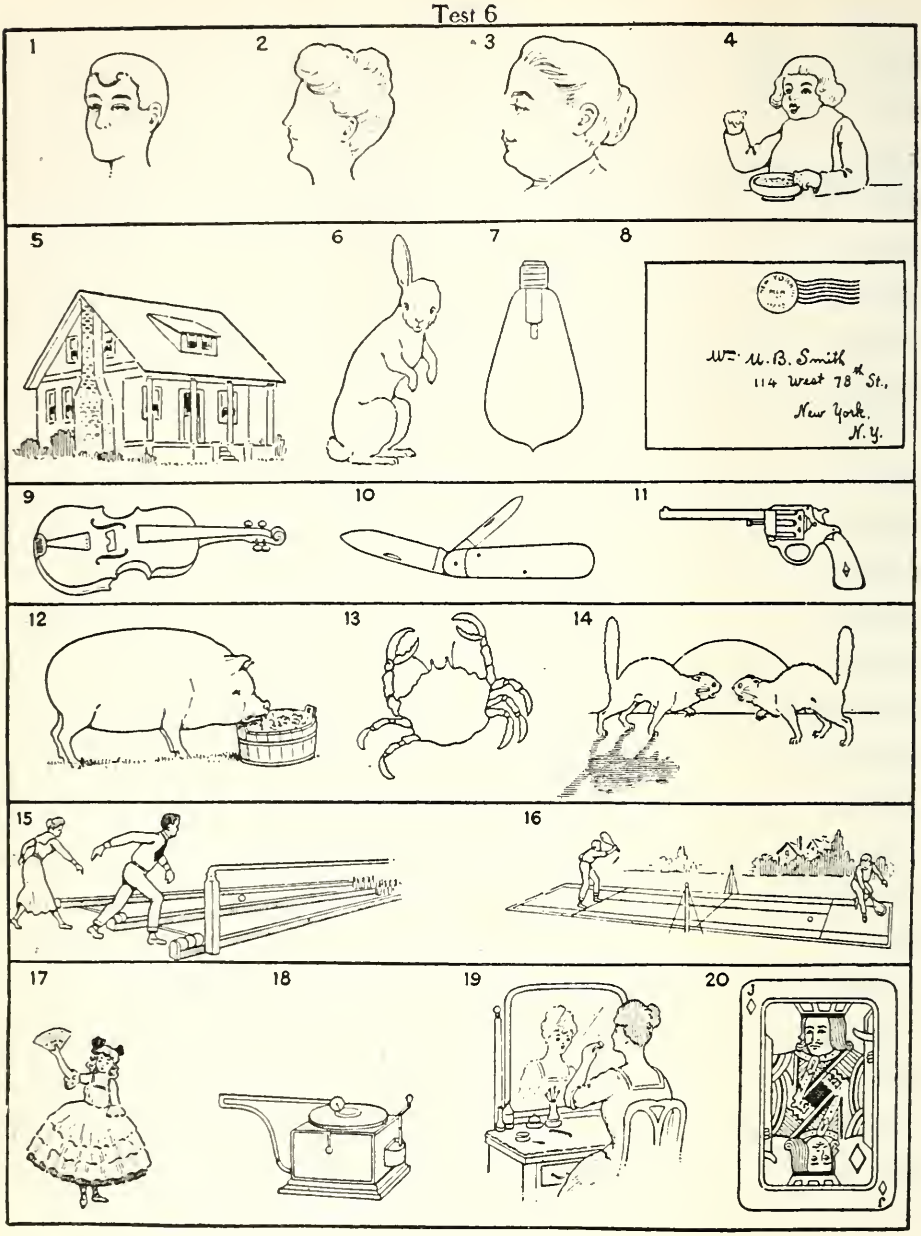
Two other important results have come from the Army testing program; one deals with occupational groups, and the other with national groups. Table 61 shows the median

Test 5

650	650	10243586	10243586
041	044	659012534	6590211354
2579	2579	388172902	381872902
3281	3281	631027594	631027594
55190	55102	2499901354	2499901534
39190	39190	2261059310	2261659310
658049	650849	2911038227	2911038227
3295017	3290517	313377752	313377752
63015991	63019991	1012938567	1012938567
39007106	39007106	7166220988	7162220988
69931087	69931087	3177628449	3177682449
251004818	251004418	468672663	468672663
299056013	299056013	9104529003	9194529003
36015992	360155992	3484657120	3484657210
3910066482	391006482	8588172556	8581722556
8510273301	8510273301	3120166671	3120166671
263136996	263136996	7611348879	76111345879
451152903	451152903	26557239164	26557239164
3259016275	3295016725	8819002341	8819002341
582039144	582039144	6571018034	6571018034
61558529	61588529	38779762514	38779765214
211915883	219915883	39008126557	39008126657
670413822	670143822	75658100398	75658100398
17198591	17198591	41181900726	41181900726
36482991	36482991	6543920817	6543920871

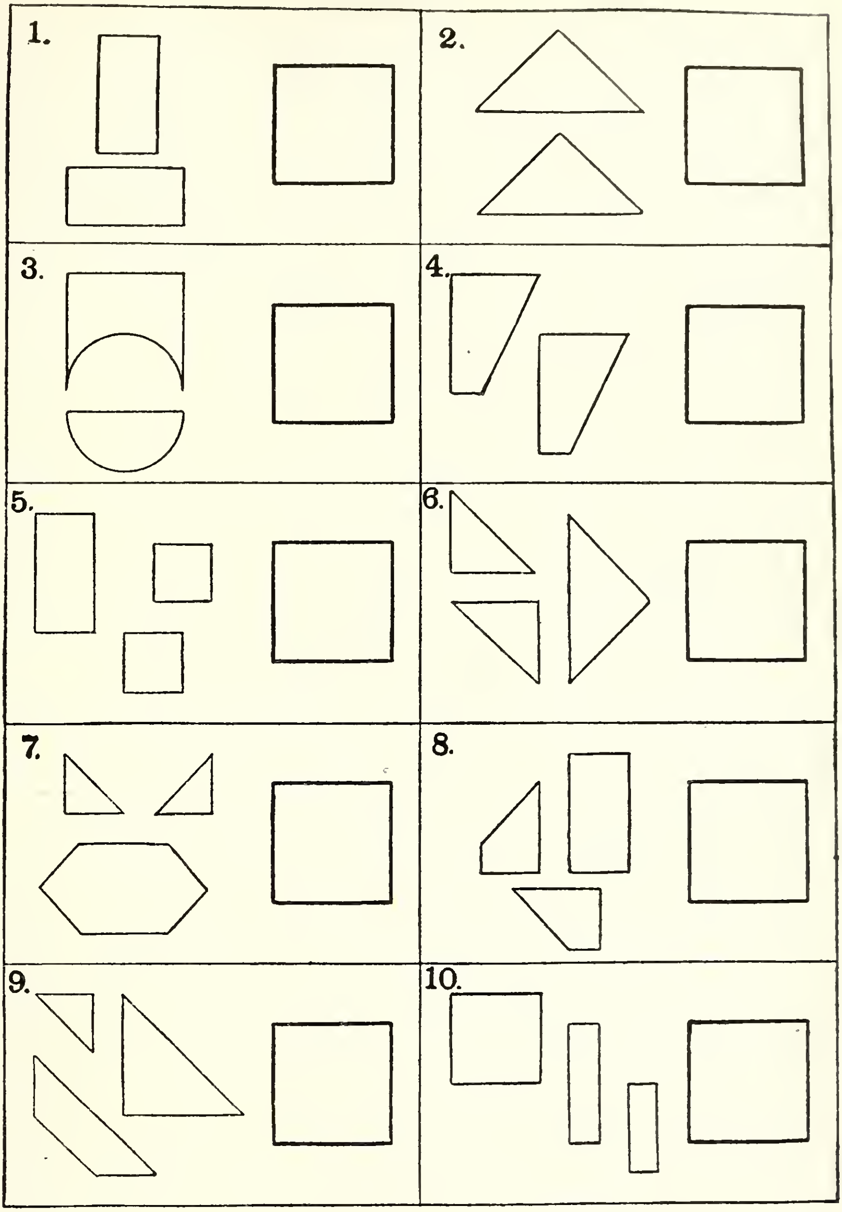
scores and the range of the middle 50 per cent for 56 occupations found in fairly large numbers in the Army. The significance of these figures for selecting and training employees is not easy to summarize. However, in many industrial situations, persons who fall in the lowest quarter of the group in Alpha Scores have been found to be poor risks

Test 6


1
2
3
4
5
6
7
8
Wm. U. B. Smith
114 West 78th St.,
New York.
N.Y.
9
10
11
12
13
14
15
16
17
18
19
20
J

unless they compensate by outstanding qualities of other sorts. In some situations Alpha Scores in the highest quarter of a group have been found to be associated with lack of interest in routine work and with short periods of employment. The United States Army psychologists also developed

Test 7



an elaborate set of Oral Trade Tests, Chapman (1921), which proved to be effective in detecting trade knowledge among recruits.

The inclusion of many foreign-born men in the United States Army in 1917–1918 led to comparisons of national groups by means of Alpha and Beta scores. Those who

TABLE 61

Occupational Intelligence Levels

Occupation	*Number of Cases*	*First Quartile*	*Median*	*Third Quartile*
Accountant and auditor	75	197	223	292
Actor	26	91	151	211
Auto repairer	189	99	146	187
Baker	40	86	146	177
Barber	58	48	87	125
Blacksmith	29	56	87	115
Bookkeeper	227	200	244	281
Brakeman, railroad	33	84	120	159
Business man	238	117	173	219
Butcher	83	68	111	156
Carpenter	359	75	127	175
Chauffeur, mechanic	260	69	118	174
Chauffeur	583	81	120	167
Chemist	52	201	253	286
Clerks	1,438	135	186	235
Conductor	78	80	132	183
Cook	228	54	93	147
Dentist	28	157	195	261
Draftsman	30	198	231	263
Druggist	70	123	174	219
Electrician	174	109	162	214
Engineer, M. E. C.	73	199	250	290
Engineer, steam	47	86	115	156
Executive	222	174	220	271
Farmer	2,017	57	97	153
Fireman, railroad	112	62	103	155
Fireman, stationary	123	52	87	159
Fisherman	49	49	79	139
Foreman	189	116	173	231
Inspector	65	105	161	205
Laborer, unskilled	2,436	47	87	140
Laborer, skilled	976	83	130	179
Lineman	58	66	115	162
Lumberman	301	67	115	166
Machinist	518	89	136	179
Mason	48	72	122	156
Mechanic	120	96	147	193
Musician	56	134	169	209
Operator and lineman	41	61	106	173
Painter	191	76	115	175
Plumber	228	87	129	168
Policeman and detective	26	111	139	204
Printer	128	118	162	212
Sailor	42	54	109	153
Salesman	564	125	172	219
Shoemaker	319	69	115	150
Station employee, railroad	52	81	151	206
Stenographer	38	197	253	279
Stock keeper	37	105	174	210
Student	161	192	246	279
Storekeeper	83	81	141	213
Tailor	34	44	85	111
Teachers	168	217	262	309
Teamster	792	57	95	142
Textile worker	349	63	103	161
Trainman, railroad	161	77	141	187
Total	15,122	—	—	—

(Based on Alpha Scores from Yerkes, Editor, 1921. By permission of the National Academy of Sciences.)

could not read English were given the Beta. Table 62 shows that the northern European groups were superior to the southern and eastern Europeans, and that these in turn were superior to American negroes. Although these figures are important for practical educational and industrial selection, they are not considered to indicate marked hereditary differences among these groups. It is generally believed that much more research is needed to establish the effects of selective migration and environment upon test scores.

TABLE 62

Mental Ages of National Groups

From United States Army Alpha and Beta Tests

	No. of Cases	*Average MA*	*SD*
England	411	14.87	2.57
Germany	308	13.88	2.43
Canada	972	13.66	2.67
Norway	611	12.98	2.47
Ireland	658	12.32	2.60
Austria	301	12.27	2.75
Greece	572	11.90	2.45
Russia	2,340	11.34	2.83
Italy	4,009	11.01	2.60
Poland	382	10.74	2.59
* U. S. Negro	18,891	10.40	2.71

(Adapted from Brigham, 1923, Table 9. By permission of the Princeton University Press.)

* From the *Memoirs*, National Academy of Science. Yerkes, editor, 1921.

Recent revisions of the Alpha by Bregman (1935) provide two forms which eliminate items which might not be known to persons not in the Army. The scores have been equated with original Alpha forms in difficulty and in retest reliability.

III. OTHER GROUP TESTS

The period since 1918 has witnessed the production of many group tests of mental abilities, some ostensibly to appraise intelligence and others to evaluate observing, reasoning, and learning in special situations. Most of these tests extended the scope of the Army tests by adapting them for use in the lower-school grades. Tests by Otis and Pressey appeared in 1918. Haggerty (1920) and Whipple published standardized tests in 1919. A group of affiliated psychologists (Whipple, G. M., *et al.*) produced the National Intelligence Test in 1920, and Terman published a group test in 1926. Usually, group tests are presented in printed form, but an oral group test has been standardized by Stump (1935) that consists of verbal and number multiple-choice items and problems. He considered that its advantages were: (*a*) allowing all students the same amount of time for each item, (*b*) allowing each student to attempt every item, and (*c*) saving the expense of printed forms. Two alternative sets of questions have been prepared for Grades IV to VIII.

Snedden (1927) experimented with a vocabulary test that was disguised in a questionnaire on those traits that may have some hereditary significance. The subject was asked which of his two parents possessed the greater amount of a trait, such as being gentle, meticulous, or sanguine. Seventy-five words of graded difficulty were selected and standardized on several hundred persons. Correlations with Stanford-Binet MA's were found to be in the neighborhood of .70. The interview form was further developed by Maizlish (1936), who issued a likes-dislikes questionnaire, and asked the subject to give reasons for his answers. The reasons showed whether or not the word was understood. This test, when given to individuals, was found to correlate .77 with Kuhlmann-Anderson tests; when given to a group, the correlation dropped to .50.

A popular variety of intelligence test for adults is known as the spiral omnibus type. This type was designed to eliminate the need for accurate timing of short periods. A test was desired which would give about the same total scores as the United States Army Alpha Test during a 20- or 30-minute period of continuous work. Hence, materials much like those of the Alpha were mixed together by rotating or spiraling among the tests. Table 63 shows a practice sheet for such a test. Among the forms now available, the Otis Self-Administering Test (1922), the Henmon-Nelson (1931) Self-Scoring Tests, the Pressey Classification Test (1922), and the O'Rourke General Classification (1932) are widely used. Some of these include more emphasis upon school achievement than others.

Thurstone (1919) inaugurated a series of mental tests for college entrance under the auspices of the American Council on Education. Annual editions have been applied to thousands of high-school graduates. It consists of types of tests, such as vocabulary, mathematics, verbal analogies, and learning an artificial language, which are administered in periods so short that few if any students complete the work. Similarly, Thorndike (1920) began a series of intelligence tests for college entrance. This includes tests of word meaning, mathematics, and special information for which 230 minutes are allowed. A similar excellent series has been designed by Toops (1937) called the Ohio State University Psychological Test, and recently two other series of this sort have been started, one by the College Entrance Board of Examiners, and the other by the Cooperative Test Service.

Another group of investigators have concentrated on nonlanguage situations which are described in Chapters 12 and 13. Items illustrating some of these tests are given below in the description of Thurstone's Primary Abilities Tests.

Six scales will be described as illustrations of particular techniques: The California Test of Mental Maturity; the McCall, Herring, Loftus Comprehensive Battery; Thorndike's CAVD Tests; the Kuhlmann-Anderson Tests; Thurstone's Test of Primary Traits; and the Bellevue Scale. All of these scales have subtests which were usually selected to have high correlations with a criterion and low correlations with each other, as in the case of the United States Army

TABLE 63

O'Rourke General Classification Test, Senior Grade

Grade of difficulty: High-school senior, College freshman

Directions and Samples

The directions and samples on this sheet are to show you the kinds of items you will find in the test. Study them carefully so you will know how you are to answer each kind of item. There will be no directions nor explanations in the test.

You are to write your answers always on the line at the right of the item.

You will be allowed 10 minutes to study the samples on the front and back of this page and answer those not answered.

WRITE ANSWERS HERE

Meaning. If the word in capital letters fits into the meaning of the sentence, write "correct" on the line at the right. If the meaning is not correct, write the number of the word which does fit. . . .

Example: The clear day was DETAILED for the picnic. (1) ORDERED (2) INVENTED (3) IDEAL (4) EXCUSED (5) DISCOVERED *3*

"DETAILED" does not fit the meaning of the sentence, but "IDEAL," which is marked "3," does, so "3" is written on the line at the right.

Relations. The first two words in each set are related in some way. Write a word which is related to the third word in the same way as the second is to the first. The word you write must begin with the letter before the answer line.

Example: SHOE is to FOOT as HAT is to . . . . H *Head*

"Head" is written on the line at the right, because it begins with "H" and a hat is worn on the head, just as a shoe is worn on the foot.

Information. Five ways of completing the statement are suggested. Write the number of the one which makes the true statement.

Example: Water is heavier than (1) paint (2) granite (3) iron (4) wood (5) sand *4*

Spelling. If every word in the sentence is correctly spelled, write "correct" on the line at the right. If you find an incorrect word, spell that word correctly on the line at the right.

Example: His education was a great advantage . . . . *advantage*

Opposite. Write the number of the word which means the opposite of the word in capital letters.

Example: The opposite of LIGHT is (1) tinted (2) small (3) bright (4) dark (5) damp *4*

"Dark," marked "4," means the opposite of "LIGHT," so "4" is written on the line at the right.

Grammar. If the sentence is grammatically correct, write "correct" on the line at the right. If you find the sentence incorrect, write what would with the least changes express the meaning correctly.

Example: The books is sold at the store. . . . . *are*

(By permission of L. J. O'Rourke, 1935 edition, and The Psychological Institute, 3506 Patterson St., N. W., Washington, D. C.)

Tests. Thurstone's battery, however, was selected to include only those tests which had low intercorrelations with all other appraisals of primary traits.

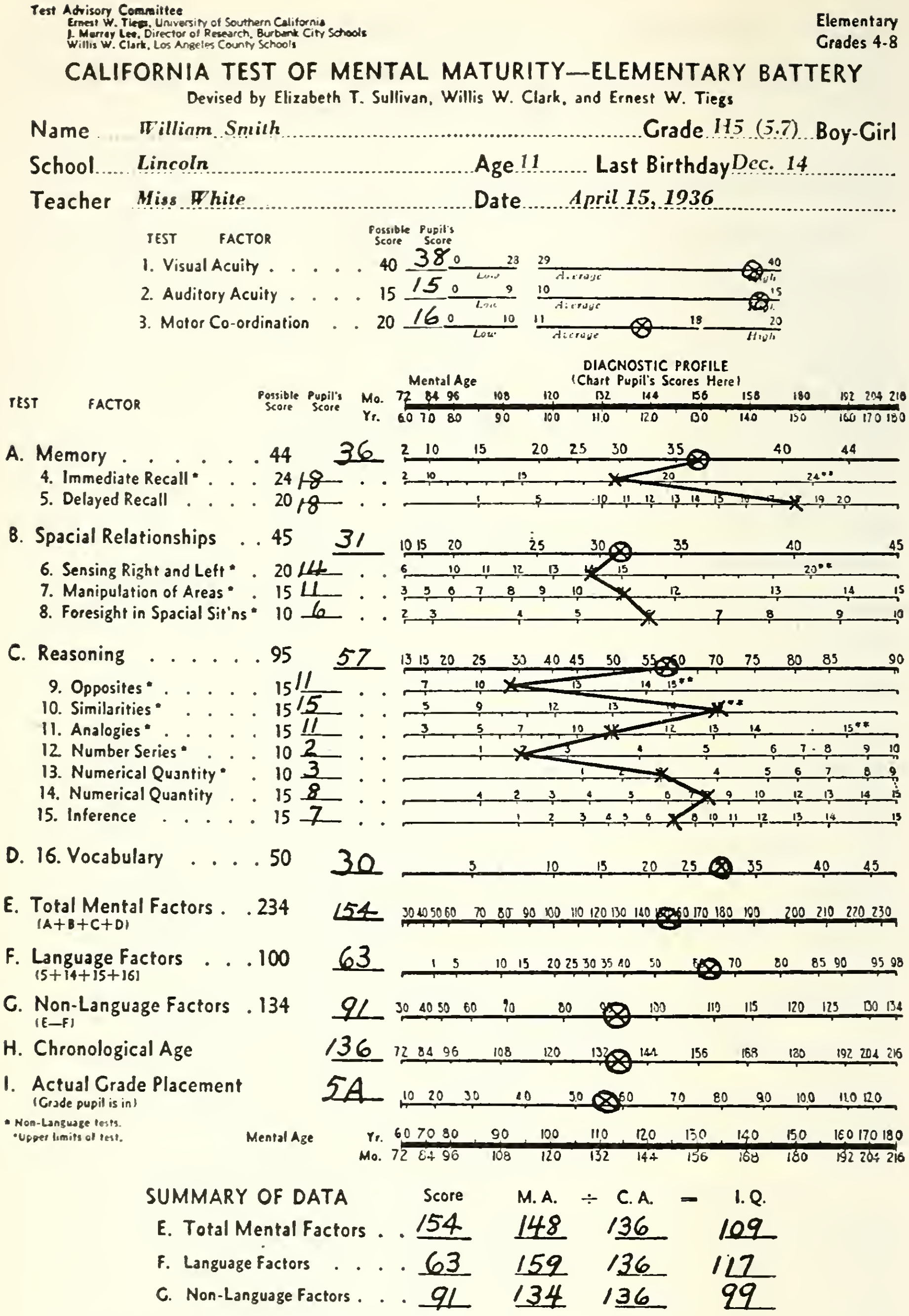
A. The California Tests of Mental Maturity

The California Tests of Mental Maturity were published by Sullivan, Clark, and Tiegs (1937) in four batteries, one for kindergarten and first grade, and the others for Grades I to III, IV to VIII, and VIII to XIV. In each battery sixteen subtests are distributed under five sections. The first section is designed to detect gross visual, hearing, and motor handicaps. The second section contains one immediate and one delayed recall test; the third, three tests involving spatial relationships; the fourth, seven tests of verbal and numerical reasoning; and the fifth, a fifty-item multiple-choice vocabulary test. The test norms allow one to draw a profile, Figure 39, showing 16 separate scores as well as total scores for verbal, nonverbal, and the whole test. The reliabilities for the subtests run from .70 to .95 and for the totals from .90 to .96. The total test requires about 90 minutes. It is one of the most extensive of its kind, and the profile of scores presents a clear picture of skills which have been found to be somewhat independent.

B. McCall, Herring, Loftus Comprehensive Battery

A comprehensive battery by McCall, Herring, and Loftus (1937) is typical of a recent development of tests for use in schools. It includes four sorts of information tests:

1. An intelligence test which consists of 150 rows each of five words or numbers (The task is to cross out the word or number that does not belong with the others. The scoring key for the words was patterned after an earlier test of McCall's which gave most credit to the answers made by the most intelligent pupils as indicated by a combination of teachers' ratings and other tests.)
2. An achievement test of knowledge in nineteen fields, such as buying and using things, arts and crafts, enjoying things, and school subjects



Test Advisory Committee
Ernest W. Tiegs, University of Southern California
J. Murray Lee, Director of Research, Burbank City Schools
Willis W. Clark, Los Angeles County Schools

Elementary
Grades 4-8

CALIFORNIA TEST OF MENTAL MATURITY—ELEMENTARY BATTERY
Devised by Elizabeth T. Sullivan, Willis W. Clark, and Ernest W. Tiegs

Name *William Smith* Grade *H5 (5.7)* Boy-Girl
School *Lincoln* Age *11* Last Birthday *Dec. 14*
Teacher *Miss White* Date *April 15, 1936*

TEST FACTOR	Possible Score	Pupil's Score
1. Visual Acuity	40	38
2. Auditory Acuity	15	15
3. Motor Co-ordination	20	16

DIAGNOSTIC PROFILE
(Chart Pupil's Scores Here)

TEST FACTOR	Possible Score	Pupil's Score	
A. Memory	44		36
4. Immediate Recall *	24	18	
5. Delayed Recall	20	18	
B. Spacial Relationships	45		31
6. Sensing Right and Left *	20	14	
7. Manipulation of Areas *	15	11	
8. Foresight in Spacial Sit'ns *	10	6	
C. Reasoning	95		57
9. Opposites *	15	11	
10. Similarities *	15	15	
11. Analogies *	15	11	
12. Number Series *	10	2	
13. Numerical Quantity *	10	3	
14. Numerical Quantity	15	8	
15. Inference	15	7	
D. 16. Vocabulary	50		30
E. Total Mental Factors (A+B+C+D)	234		154
F. Language Factors (5+14+15+16)	100		63
G. Non-Language Factors (E—F)	134		91
H. Chronological Age			136
I. Actual Grade Placement (Grade pupil is in)			5A

* Non-Language tests.
**Upper limits of test.

SUMMARY OF DATA

	Score	M. A. ÷	C. A. =	I. Q.
E. Total Mental Factors	154	148	136	109
F. Language Factors	63	159	136	117
G. Non-Language Factors	91	134	136	99

0 Low 28 29 Average 40 High
0 Low 9 10 Average 15 High
0 Low 10 11 Average 18 20 High
Mental Age
Mo. 72 84 96 108 120 132 144 156 168 180 192 204 216
Yr. 6.0 7.0 8.0 9.0 10.0 11.0 12.0 13.0 14.0 15.0 16.0 17.0 18.0
2 10 15 20 25 30 35 40 44
2 10 15 20 24**
1 5 10 11 12 13 14 15 16 17 18 19 20
10 15 20 25 30 35 40 45
6 10 11 12 13 14 15 20**
3 5 6 7 8 9 10 12 13 14 15
2 3 4 5 7 8 9 10
13 15 20 25 30 40 45 50 55 60 70 75 80 85 90
7 10 13 14 15**
5 9 12 13 14 15**
3 5 7 10 12 13 14 15**
1 3 4 5 6 7 8 9 10
1 2 4 5 6 7 8 9
4 2 3 4 5 6 7 9 10 12 13 14 15
1 2 3 4 5 6 8 10 11 12 13 14 15
5 10 15 20 25 35 40 45
30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 230
1 5 10 15 20 25 30 35 40 50 60 70 80 85 90 95 98
30 40 50 60 70 80 90 100 110 115 120 125 130 134
72 84 96 108 120 132 144 156 168 180 192 204 216
10 20 30 40 50 60 70 80 90 100 110 120
Mental Age
Yr. 6.0 7.0 8.0 9.0 10.0 11.0 12.0 13.0 14.0 15.0 16.0 17.0 18.0
Mo. 72 84 96 108 120 132 144 156 168 180 192 204 216

FIGURE 39. California Mental Maturity Test Profile

(Copyright, 1937, by E. T. Sullivan, W. W. Clark, and E. W. Tiegs. By permission of the Southern California School Book Depository, Los Angeles, California, and the authors.)

3. An educational background questionnaire
4. A school practices questionnaire which evaluates the pupil's knowledge and opinion of his home, community, and school as cultural and democratic institutions

This is a pioneer work which has been standardized on about 20,000 cases. The reliability reported for the intelligence test is .97; for the achievement, .96; and the questionnaires, .91. The four scores provide material which is much needed for appraising a child's progress in fields which are important.

C. Thorndike's CAVD Scales

This series of tests is distinguished by its continuity, the thoroughness of its scaling, and its use in Thorndike's analysis of mental ability. The selection of items was based on three sorts of criteria (p. 63):

> (1) Of psychological theory: – (*a*) that responding to parts, elements, or aspects of situations is more "intellectual" than responding to gross total situations; (*b*) that responding to parts, elements, or aspects which do not present themselves separately to sense but must be abstracted is more intellectual than responding to those which do; (*c*) that responding to relations between objects is more intellectual than responding to objects; (*d*) that, in particular, responding to so-called subjective or logical relations, such as likeness and difference, is more intellectual than responding to the so-called objective relations of space and time; (*e*) that organizing several mental connections or habits to secure certain results, "thinking things together," as James put it, is more intellectual than using one habit at a time; (*f*) that responses to novel situations are likely to be more "intellectual" than responses to familiar situations.
>
> (2) Of the theory of measurement: – (*a*) that the tasks representing any one ability should be capable of very fine gradation from very easy to very hard; (*b*) that they should be capable of very wide extension by alternates at any degree of difficulty; (*c*) that, so far as possible, any one ability should represent in some real and useful sense something varying only in amount, so that the different degrees of it might properly be represented by numbers.

(3) Of common sense: – (*a*) that the tasks should be from among those which had high standing on the basis of correlations with reasonable criteria; (*b*) that they should be convenient for use in the actual measurement of intellect; (*c*) that they should be tasks concerning which subjects for experiment were obtainable.

No attempt was made to select items which might be particularly adapted to the measurement of original nature. The tests were limited in most cases to pictorial material or verbal responses to questions. In order to simplify research, responses to persons and to objects were not included. The scale consists of four subtests whose initials are C, A, V, and D.

C. *Completion Test.* This requires a person to supply words to make statements true and sensible. These statements range from easy items suitable for 3-year-olds, such as:

"At night you sleep in a ______________."

to the hardest level for adults:

"It must ______________ seem to the wisest ______________ men, when brought into contact with the great things of nature that ______________ they ______________ is ______________ nothing ______________ to the infinitude of ______________ they are ignorant."

A. *Arithmetic Problems.* These include concepts of number and size at the easiest level, such as counting 3 pennies, and choosing the littlest of 3 squares, and at the hardest level:

"Let m_1, m_2, m_3, etc., be any numbers.
Let n be their number, that is, n tells how many *m's* there are.
Let S (   ) mean the sum of.
What name will you give to $S \frac{(m's)}{N}$?"

V. *Vocabulary Tests.* These require definitions of single words. The easiest levels exhibit a row of five pictures and require a correct response to such directions as, "Put your finger on the horse." The hardest levels are printed multiple-choice tests. Thus, "auricular: 1 golden 2 heard 3 jointed 4 distinct 5 clear."

D. *Direction Tests.* These are tests of one's understanding of connected discourse. At the easiest levels the child is told to, "Put your hands behind you" or "Turn the box upside down." At the hardest levels a long paragraph is to be read and then questions about the meaning are to be answered.

The construction of the scale involved a preliminary tryout of a large number of items on fairly large age groups. The items were then arranged in order of difficulty and grouped into 17 levels of difficulty. The lowest level is suitable for 3-year-olds and the highest for superior adults. The levels are separated by steps of approximately equal difficulty. Finally a selection was made so that each level contains 40 items, 10 in each subtest. Each correct answer was given one point credit. The lowest 5 levels were designed for individual testing, and the higher levels for group testing. There are no time limits. The pupils are started at a level where they can answer all the items correctly, and they continue to a level where they fail practically everything.

The odd-even reliabilities of scores on the subtests are very high, and the subtests also show high intercorrelations. Three scores can be obtained from this material:

1. *Level of achievement or altitude,* which is indicated by the level at which a person succeeds in one-half the items
2. *Width* at a particular level, which is the per cent of items passed at a given level
3. *Area,* which is the sum of all successes at all levels

Thorndike's studies furnished evidence that level of ability increased very slowly during adolescence, but that width increased rapidly during this period. Hence, it is important to know whether a test score is dependent largely upon altitude or breadth of items. Figure 40 shows a theoretical diagram of a person's intellectual growth. The lowest level has little or no breadth or variety of response, whereas the highest levels have great variety. Thorndike points out that unless a scale is carefully constructed, it may depend more on breadth than altitude, but no clear interpretation is possible. Figure 40 also emphasizes two aspects of measurement. One need is that for measuring independent abilities separately, for one may show great altitude in some fields and not in others. Another need is that of a true zero point. If Thorndike's hypothesis is correct, a zero of ability can be estimated as the level at which there is no breadth or variation. This hypothesis is considered in Chapter 22.

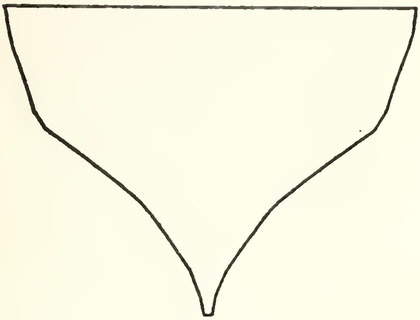


FIGURE 40

The Area, Altitude and Breadth of Mental Ability

(From Thorndike, 1926, p. 474. By permission of the Bureau of Publications, Teachers College, Columbia University.)

Thorndike *et al.* (1935) published four new forms of the CAVD tests for high school and college levels. The scoring has been made quicker, easier, and more rigid than in the previous forms.

D. Kuhlmann-Anderson (1927) Intelligence Tests

This is a battery of tests which were developed over a ten-year period by the Research Division of the Minnesota State Board of Control. The present scale includes 35 tests which have survived out of more than 100 which were tried out. The tests were selected because they gave the largest and most consistent increases in scores for successive age levels. No attempt was made to secure internal consistency or subtest reliability because Kuhlmann believed that such procedures could not improve the scale. Total scores have been found to be as reliable as most group tests of this kind.

The tests are arranged in nine booklets for various grade groups from the 1st to the 12th grade. Each booklet includes 10 tests which are arranged in order of difficulty. Booklets for adjacent grade levels overlap considerably by having some tests in common.

Of the 35 tests, 3 make use of pictures, 6 of geometric patterns, 8 of mathematics, 2 of new associations, and 16 of verbal relations and information. The test for the earlier ages have more pictorial and less number and reading material than those for the later ages. All the tests have time limits. At the earlier ages these are liberal enough for 95 per cent of the group to attempt each problem, but at the later ages speed during 1½- or 2-minute periods is very important.

Each test for the lower grades consists of from 5 to 10 items, and each at the upper levels has approximately 15 items. The mental age equivalent for each score of each test is given in a table. A person's mental age for each test is recorded to give a profile of his highest and lowest abilities, and a median mental age is calculated to represent general intelligence.

E. Thurstone's Test of Primary Traits

Although nearly all authors of tests have endeavored to analyze their tests subjectively, two authors, Spearman and Thurstone, have worked very hard to devise quantitative methods of showing what basic patterns were important for success in a particular test. Chapter 7 describes Spearman's work. Recently Thurstone has completed monumental researches which indicate that certain basic patterns appear in the populations tested. Thurstone presented a group of college students with 57 carefully controlled tests during 15 hours of work. The test included nearly every variety of verbal and nonverbal form of test extant. They did not include adjustments to persons or performance tests. The scores for each test were correlated with each of the others, and the resulting matrix yielded 12 factors, of which 9 were sufficiently large to be identified. The tests which had large loadings, and hence are considered to be the best indicators of factors, are shown in Table 64.

Using these and similar tests, a scale for superior adults has been constructed which discloses a person's relative ability in these primary traits. A similar scale for elemen-

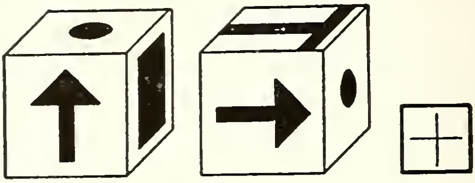
TABLE 64

Tests of Primary Abilities

1. Spatial Relations (S)

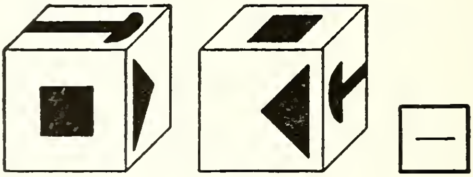
Cubes (adapted from Brigham, 1932): factor loading, .626
The drawings in this test represent cubes. There is a different design on each face of the cube. A cube has six faces.

Notice that both of the drawings below can represent the same cube. Be sure you see that the first and second drawings represent the same cube turned into two different positions. Since both drawings can represent the same cube, a plus sign (+) has been placed in the blank square at the right.



1A

Notice that the two drawings below represent two different cubes and that a minus sign (−) has been placed in the blank square at the right. Be sure that you see that it would be impossible to turn the cube shown in the first drawing so that it would look EXACTLY like the cube shown in the second drawing. Unless you see this clearly, you cannot solve the test items. There is a different design on each face of the cube.



1B

2. Perceptual Speed, Visual (P)

Word grouping: factor loading, .573
In the line below, notice that the four words, dog, lion, cat and giraffe, can be grouped together because they are all names of animals. The word chair does not belong with the others because it is not the name of an animal. Since chair is the second word, 2 is written in the blank at the right.

1—dog 2—chair 3—cat 4—lion 5—giraffe 2

Similarly, four of the words in the line below can be grouped together because they are alike in some way, while one of the words does not belong because it is different. Write the number of that word in the blank at the right.

1—carrot 2—radish 3—beet 4—book 5—turnip ___

3. Number Calculations (N)

Addition: factor loading, .755
Addition is a simple test of ordinary addition of seven two-digit numbers.

$$\begin{array}{r} 17 \\ 64 \\ 93 \\ 21 \\ 14 \\ 17 \\ \underline{65} \end{array}$$

Multiplication: factor loading, .812
Multiplication is a simple test involving the multiplication of six-digit numbers by a single-digit number.

$$\begin{array}{r} 7245986 \\ \underline{4} \end{array}$$

TABLE 64—*Continued*

4. Verbal Relations (V)

Inventive Opposites: factor loading, .635

This is a test of ability to think of words. Think of two different words opposite in meaning to the word narrow below. One word should begin with b. The other should begin with w. The words are broad and wide. These words have been written in the blanks.

narrow b______________ w______________

Now think of two words opposite in meaning to the word large. The first should begin with l; the second with s.

large l______________ s______________

The words are little and small. Write little in the first blank. Write small in the second.

strong f______________ w______________
wrong r______________ c______________
dark b______________ l______________

5. Word Forms (W)

Anagrams: factor loading, 5.34

Make as many different words as you can, using only the letters in the word G–E–N–E–R–A–T–I–O–N–S. You may use long or short words and may include the names of persons, places, or foreign words. In any one word do not use a letter more times than it appears in G–E–N–E–R–A–T–I–O–N–S.

Sample words have been written in the first few lines. Continue writing as many words as you can, using only the letters given.

G–E–N–E–R–A–T–I–O–N–S

1. ART
2. ERA
3. SNORE
4. ______________

Disarranged Words: factor loading, .512

Rearrange the letters on each of the following lines to spell the name of an animal. In the first line, the letters (ebar) can be arranged to spell bear, which is written in the blank space. In the next line, the letters (odg) spell dog, which is written in the blank space. In the same way the letters (atc) spell cat.

ANIMALS

ebar	bear
odg	dog
atc	cat

Rearrange the letters on each of the following lines to spell the name of a boy. The first two names have already been written for you. Write the third.

BOYS' NAMES

lpau	Paul
rcla	Carl
honj	______

Rearrange the letters on each of the following lines to spell the name of a bird.

BIRDS

uckd	__________
cowr	__________
wahk	__________

6. Immediate Memory Span (M)

Word-number: factor loading, .529

Word-number was prepared as a test involving memorizing. The subject memorizes a set of paired associates. Each stimulus word is to be associated with a response number. In the recall the subject is given the stimulus word, and he is asked to write the corresponding response number. The test is arranged with instructions and a fore-exercise followed by a recall. A second fore-exercise, which is longer, is then given.

TABLE 64—*Continued*

It is followed by a recall. The test proper with twenty words and associated numbers is then presented. This is followed by the recall. The test really consists of three sections with a presentation and recall in each section. This was done in order to make sure that the subjects understood the nature of the task.

Number-number: factor loading, .664

Number-number is a paired-associates test in the same form as the two previous tests in which the stimulus consists of a two-digit number, and the response is another two-digit number. This test is also given in two sections with two parts for each section. Together with the instructions the subject is given five paired numbers to associate. He is then asked to recall the five response numbers when the stimulus numbers are presented. He is given the opportunity to write the numbers if he wants to learn them in that manner. Then follow the memorizing of the twenty pairs of numbers and the recall in which the twenty stimulus numbers are presented in random order. The subject is asked to fill in the response numbers.

7. Induction (I)

Number series: factor loading, .503

The numbers in each row of this test follow one another according to some rule. You are to find the rule and fill in the blanks to fit the rule.

In the example below each number can be obtained from the one before it by the rule add 2. The blanks have been filled in accordingly.

2 4 6 8 10 12 14

Find the rule in the series below and fill in the blanks. You may use addition, subtraction, multiplication, division, or any combination of these.

10 8 11 ___ 12 ___ 13

The above series goes by alternate steps of subtracting 2 and adding 3. You should have written 9 and 10 in the blanks.

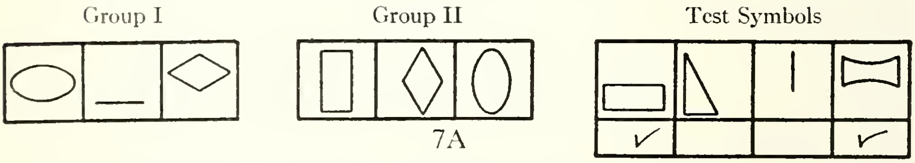
Find the rule in each series below and write the numbers in the blanks accordingly. There is a different rule for each line. Go right ahead. Do not wait for any signal.

19 18 17 ___ 15 14 ___
8 11 14 ___ 20 ___
27 ___ 23 23 19 19

Figure Classification: factor loading, .405 (Also called Spearman's Form Analogies Test)

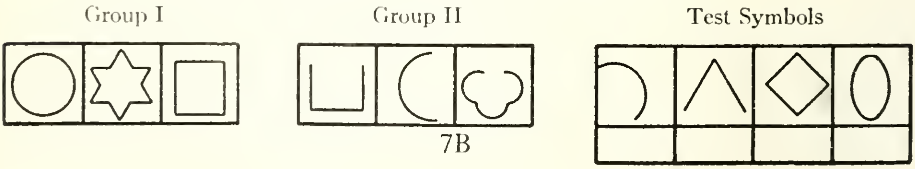
In each line below, there is a rule by which the symbols in Group I differ from those in Group II. There is a new rule for each line. Your problem is to discover the rule in each line. Some sample problems are worked for you below.

In the first line below the rule is that the symbols in Group I are horizontal while those in Group II are vertical. Each of the test symbols at the right belongs either to Group I or to Group II. The test symbols that belong to Group I have been checked.



7A

The rule in the problem below is that the figures in Group I are closed while those in Group II are open. Now check the test symbols that belong to Group I.



7B

TABLE 64—*Continued*

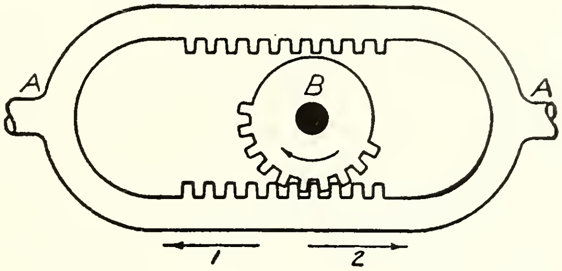
You should have checked the third and fourth symbols. They are closed figures.

8. Reasoning (R)

Arithmetical Reasoning: factor loading, .583
Arithmetical Reasoning contains nineteen problems and is similar to current tests of this type.
Mechanical Movements: factor loading, .414
In this test you will be shown pictures of mechanical movements. You will be asked questions about them.

In each picture the part that makes the others move is called the driver. The solid black circles represent axles which can turn but cannot move from where they are shown.

Now answer the questions after each of the pictures below. Go right ahead.



8A

1. If B starts moving in the direction shown, which way will A move, 1 or 2? . . . . ______
2. In which direction will A be moving when B has turned half way around from where it is now? . . . . ______

Vocabulary, general: factor loading, .545
Word Knowledge is the Thorndike Vocabulary Test.

9. Deduction (D)

False Premises: factor loading, .578
This is a test of your ability to tell the difference between good and bad reasoning. You must judge only the reasoning in the following arguments because every statement is false or even absurd.

The first argument below is good reasoning and is marked plus (+). The second argument appears similar but is bad reasoning and is marked minus (−).

All haystacks are catfish. All catfish are typewriters.
Therefore all haystacks are typewriters. +

All haystacks are typewriters. All catfish are typewriters.
Therefore all haystacks are catfish. −

Reasoning: factor loading, .525
This is a test of your ability to tell the difference between good and bad reasoning.

The first argument below is good reasoning and is marked plus (+). The second argument appears similar but is bad reasoning and is marked minus (−).

All sports are dangerous, and football is a sport.
Therefore, football is dangerous. +

Some sports are dangerous, and football is a sport.
Therefore, football is dangerous. −

Now mark the two arguments below in the same way.

All wealthy men pay taxes. Mr. White pays taxes.
Therefore, Mr. White is wealthy. ______

All wealthy men pay taxes. Mr. White is wealthy.
Therefore, Mr. White pays taxes. ______

The first argument above is bad and should have been marked minus (−). The second should have been marked plus (+).

(Arranged from Thurstone, 1938. By permission of the Editor, *Psychometrika Monographs.*)

tary-school groups is being constructed. The results of applying such scales are not easy to predict. Doubtless, total scores will correlate with other intelligence test scores fairly well, so that no marked new contribution will be seen when total scores are used. The scores from the subtests, however, are likely to be more predictive of success when success is dependent upon particular combinations of special abilities.

Table 65 shows the factor loadings for the tests just described together with the per cent of the total variance, which is indicated in the h^2 column, and the retest reliability. From this table it appears that some tests have fairly heavy loadings in more than one factor. Thus, Figure Classification shows a loading of .39 in Factor S, *Spatial Visualization;* .40 in Factor I, *Induction;* and .40 in Factor D, *Deduction.* Such tests are not of as great analytical value as *Multiplication,* which is well accounted for by Factor N, *Number Facility.* It is also clear from the h^2 column that some of the items are not very completely accounted for by these factor loadings. Since the tests are shown to have high reliabilities, it is probable that tests with small h^2 values, such as Figure Classification, demand skills or adjustments not well appraised by this analysis. This condition can only be remedied by more extensive study.

F. The Bellevue Scale

Although this scale is designed for individual administration, it is placed in this chapter because its form and derivation closely resemble the group tests which have been discussed. The Bellevue Scale is an example of a combination of techniques taken from Binet-type group intelligence and performance tests. Published by Wechsler in 1939, it consists of a standardization of ten widely used types of tests. Five of the tests, termed verbal, may be used for those with defective vision, and five, termed performance tests, may be applied to those with language or hearing handicaps.

TABLE 65

Factorial Loadings of Mental Tests

Factor	*Test*					*Factors*					h^2	*Reliability*
		S	P	N	V	M	W	I	R	D		
Spatial (S)	Cubes	.63	.21	.20	.06	.03	.15	.10	.29	.27	.75	.86
Perceptual Speed (P)	Word Group	.11	.57	.08	.46	.01	.13	.04	.23	.17	.71	.95
Number (N)	Addition	.06	.03	.75	.19	.05	.04	.02	.04	.22	.68	.85
	Multiplication	.02	.09	.81	.02	.07	.28	.07	.14	.03	.79	.73
Verbal Relations (V)	Inventive Opposites	.07	.19	.05	.63	.22	.33	.03	.31	.05	.81	.91
Word Forms (W)	Dis. Words	.06	.22	.06	.10	.10	.51	.03	.22	.22	.55	.95
	Anagrams	.04	.00	.09	.18	.13	.53	.14	.10	.20	.56	—
Memory Span (M)	Word-Numb.	.01	.07	.02	.12	.53	.06	.03	.09	.22	.42	.91
	Numb.-Numb.	.04	.08	.24	.02	.66	.03	.30	.06	.11	.62	.68
Induction (I)	Fig. Class	.39	.19	.08	.05	.05	.06	.40	.09	.40	.64	.90
	Numb. Series	.06	.09	.35	.30	.26	.00	.50	.09	.29	.68	.80
Reasoning (R)	Arith. Reas.	.08	.01	.38	.17	.28	.03	.33	.58	.08	.84	.69
	Vocab.	.03	.41	.12	.38	.12	.41	.22	.54	.16	.96	.84
	Mech. Mov.	.07	.19	.09	.03	.04	.02	.17	.41	.40	.67	.91
Deduction (D)	Reason	.05	.05	.04	.42	.18	.17	.20	.36	.52	.81	.82
	False Prem.	.02	.01	.03	.42	.26	.16	.19	.36	.58	.77	.50

(Arranged from Thurstone, 1938, from Tables 3 and 4. By permission of the Editor, *Psychometrika Monographs.*)

All of the tests have been standardized on groups of normal persons.

The verbal tests, all of which use oral directions and answers, include:

1. Information: 25 questions similar to the United States Army Alpha, No. 8
2. Comprehension: 12 questions like the United States Army Alpha, No. 3, such as, "What would you do if you found a letter that was already sealed, stamped, and addressed?"
3. Arithmetical Reasoning: 10 problems, to be done mentally, similar to the United States Army Alpha, No. 2
4. Digits Forward and Backward: the score is the longest span to be repeated correctly
5. Similarities: 12 pairs of words, such as "Egg – seed," for which one common aspect must be mentioned

The performance tests include:

6. Picture Completion: 15 cards, similar to the United States Army Beta, No. 6
7. Picture Arrangement: 6 series of cards to be arranged in temporal order
8. Object Assembly: 3 pictures cut apart, a manikin, a feature profile similar to those used by Pintner and Paterson (1917), and a hand
9. Block Designs: 7 designs similar to the Kohs (1927) Series
10. Digit Symbol: similar to the United States Army Beta, No. 4 (This is a paper and pencil test.)

A 42-word vocabulary test ranging in difficulty from *apple* to *traduce* is used as an alternate. The scoring of this scale, like that of many individual scales, is rather intricate. The raw scores for each item are zero, half, or full credit, based on the examiner's judgment of the correctness and completeness of oral responses or performances. The twenty-two pages of criteria for scoring leave many close decisions to the examiner's judgment. The raw scores for each of the ten tests are given equal weight, and all are converted by a table into points on a scale for which the mean is arbitrarily set at 10, and the SD at 3 for a group of 495 adults of 20 to 34 years of age. The sum of the five verbal test scores yields a verbal MA and IQ, and the five performance tests yield similar indices. A total MA and an IQ are also available. All of these indices are secured from tables which have been prepared so that the mean IQ will be 100 and the PE, 10. Since an SD is 1.4826 PE in a normal curve, the SD is approximately 15 for these IQ's. This is slightly smaller than the SD of 17 used by Terman and Merrill (1937). Reliabilities for the separate tests were not published, but the total split-half reliability corrected for attenuation was .90, and the correlations between verbal and performance sections corrected for attenuation was .83 in a group of 355 adults. The uncorrected correlations are not given. The median correlation of separate tests with total scores was approximately .65 for this same group, with *Similarities* and *Block Designs* showing the highest correlations, and *Object Assembly* and *Digit Span* the lowest. The Bellevue Scale is apparently one of the best oral scales for the examination of average adults; for the standardization has been as adequate as that of any individual test, and the test items have been selected to appeal to adults and to sample their abilities. Further research is needed to establish the optimum length of the subtests of a battery of this sort. The Bellevue Scale subtests may be too short and too subject to practice effects to be reliable enough for individual profiles to use in studies of deterioration made through a comparison of subtest scores.

IV. CORRELATIONAL ANALYSIS

Of the large number of factor analyses of cognitive processes, two sorts will be described. One group of studies attempts to discover unitary traits, and the other analyzes the effects of variations in difficulty on obtained factors.

A. Unitary Traits

It is interesting to note that no large general or *g* factor emerged from Thurstone's analysis, but separate factors appear for comparison of form, words, meaning, geometric shapes, and numbers. Moreover, comparison on a perceptual level was unrelated to comparison on a logical level, and immediate results of memorizing were quite different from other sorts of recall.

Recently Holzinger and Harman (1938) analyzed the identical matrix, provided by Thurstone, using a bi-factor method derived from Spearman's hypothesis. They found that the correlation matrix could be resolved into one general factor, which they called u_1, and nine group factors. Six of these group factors have a very high correspondence with the group factors found by Thurstone from the same data — namely, *Number, Visualizing of Space, Memorizing,* the two verbal factors, and *Deduction.* The correspondence is less close but still marked in the case of *Perceptual Speed,* and no factor appears to correspond with *Induction.* Holzinger's analysis produced two other factors which he called *Analogies* and *Rhythms.*

The main difference between these two analyses is the appearance of a large *g* or general factor in Holzinger's work but not in Thurstone's. Holzinger writes: "The presence of this factor is due to our hypothesis of its existence and the essentially positive correlation throughout. . . . It can be shown that each of the (Thurstone) group factors . . . can be expressed as a linear function of the corresponding group factor and the general factor of the Bi-factorial analysis."

Alexander (1935) secured 15 test scores and 5 grades in school on a group of 100 trade-school boys, and then calculated a factorial analysis using Thurstone's (1935) method with some variations. He concluded that five factors were needed to account for all the correlations found between these variables. A general reasoning and comparison factor, called "g," was found to correlate most highly with Number Tests, Spearman's Form Analogies Tests, and Verbal Tests in that order. A verbal factor (V) was found to be large in four verbal tests and grades in Science and English. A third factor was found principally in the block performance and mechanical interpretation tests. Two other factors were isolated. One was large in school success, particularly technical achievement in Mechanical Drawing, Mathematics, and Science. The other was largest in Shop Achievement, Kohs Block Designs, and Spearman's Dot-Pattern Test. This study is particularly interesting because it reports an important factor found in school achievement but not in standard tests. The results are also interesting because they differ from Thurstone's (1938) analysis, although using some of the same tests. Alexander's work indicates that the amount of change in factor loadings between two analyses of a test depend upon (*a*) differences in abilities of the groups tested, and (*b*) differences in the test batteries subjected to analysis. Analyses based on large batteries of tests given to large populations will undoubtedly appear and be accepted as standard.

Holzinger (1934 a.b.c. and 1935 a.b.) introduced a bi-factor method of analysis. He applied this to 78 tests which were completed by 118 children whose mean chronological age was 13.2 years and SD, 1.5 years; and also 63 tests using 100 adolescents whose mean age was 17 years, and SD, 1.9 years. He found independent factors which he called Spearman's perceptual *g* factor, verbality, mental speed, motor speed, mechanical ability, attention, oscillation, and imagination. He calculated the proportion of the variance of each test which is due to each factor and also to specific factors. Holzinger and Swineford (1936 a.b.) described and illustrated a new method of the allocation of tests for the determination of group factors and the estimation of pupils' profiles. For precise individual measurement they believe that 300 persons must be measured on 50 tests at least. A recent report by McNemar (1939) indicates that samples of 200 are not large enough to give accurate factor loadings on more than one of the factors when a battery of nine tests was used.

B. The Effects of Difficulty on Obtained Factors

A factorial analysis of the effects of "difficulty" upon test interrelations was made by Hertzman (1936). He used seven tests, three of which involved geometric figures, and four, numbers. He divided the items of seven tests into hard and easy halves, using the relative success of 186 upper classmen as his criterion of difficulty. In general, the mean score of the more difficult (D) tests was about 60 per cent of the easy (E) tests. He calculated correlations between each level and between the two kinds of material used, numerical and spatial. His results show that the easy level of any test correlated more highly with the hard level than it did with any other measure with one exception, spatial analogies. Easy spatial analogies correlated most highly with the other easy tests, and hard spatial analogies with the other hard tests. The difficult levels had a higher mean correlation (.39) than the easy levels (.33). His analysis, using both a tetrad-difference method and an iterative-factor method, yielded two factors. He also found that the average correlation between the numerical and spatial tests (.374) is greater than the average correlation among the four easy tests which involve numbers. He concluded from this and other evidence that the easy and difficult levels cannot be thought of as measuring the same ability. He believed one should hesitate to call these four tests numerical tests since the knowledge of numbers is probably not the most important element in success in all the tests.

Two important conclusions arise from this study. One is that where items in a test vary in difficulty, they may require different combinations of factors for success. The other is that the use of particular symbols in a group of tests does not signify that the main difficulties of those tests are principally dependent upon the knowledge of those symbols. The principal components may be logical patterns which are independent of the symbols used.

C. Discussion of Analysis

These studies of mental factors must be regarded as preliminary to more extensive and thorough work. Nevertheless, from relatively small batteries of tests used upon small populations, some remarkably consistent results have appeared. Among the individual workers, Thurstone, Holzinger, Spearman, and Kelley are the best known. They all find evidences of factors which correspond to visualization, perceptual speed, and simple number calculation. A general verbal factor found by authors using batteries of a dozen or so tests seems to break down, when many tests are used, into two factors, one of which involves the meaning of words and the other the form of words. Similarly, a general reasoning factor seems to be divided by more detailed analysis into induction, deduction, and problem-solving activities. Induction and deduction are, at least after practice, fairly well-standardized methods of comparison, and problem-solving involves a puzzle situation where one has to size up a somewhat novel situation and try out various hypotheses.

It seems probable from the nature of factorial analysis that the general factors which have been described are the most complex common factors found in the behavior of a particular group during a particular test series. This statement is true because factorial analyses usually seek the smallest number of factors which will account for the correlations between tests. When the tests are constructed so that the same scores may represent very different sorts of abilities, then no clear explanation of factors can be made. When tests are constructed to allow only one method for success, however, then factor analyses will yield results that can be clearly recognized by psychologists.

In conclusion, it seems probable from these facts and from facts given elsewhere in this book that the broad, general factors, such as verbal, spatial, number, and reasoning, are convenient classifications of behavior just as convenient classifications of objects are machinery, crockery, and furniture. By still further discrimination, types of machinery can be found, some of whose dimensions will have a zero correlation with each other, and hence these dimensions can be considered evidences of independent factors among machines. Likewise, a careful discrimination among verbal skills has disclosed a number of subfactors which seem to be independent when contrasted with each other, although they were found together in rough general classifications or in complex tests. We may therefore expect from further factorial analyses many valuable data on the relationships among various types of behavior from the point of view of broad general classes and of small, more precise subdivisions of classes.

12

PERFORMANCE, MECHANICAL, AND MOTOR TESTS

I. INTRODUCTION

Measurement of human ability inevitably involves spatial adjustments, either of perception or of movement, and in this division of behavior three groups of tests can be distinguished. In the first group, verbal memory or reasoning are the dominating activities, and spatial factors are probably of so little importance that they can be ignored. These tests have been discussed in Chapter 11. In the second group, memory, perception, and reasoning tests which involve form are of major importance. The second group are called *performance tests* when they are designed to appraise general intelligence, and *mechanical tests* when they are constructed to indicate specific achievement in mechanical skills. In the third group, motor skills are of prime importance, such as steadiness, speed, accuracy, strength, endurance, and complexity of co-ordination. Batteries of tests which include performance tests are described in Chapters 8 and 11. This chapter supplements them and also gives additional data on mechanical and motor tests.

II. PERFORMANCE TESTS

The difficulty of applying language tests of intelligence to various groups was apparent at an early date. Differences in language training, in speech ability, and in spoken language were so commonly found that nonlanguage tests of intelligence were among the earliest to be designed and standardized. Seguin (1846) devised a form board which was essentially the same as that now used in several scales (see Figure 42). In 1911 Witmer described a cylinder test and a revision of the Seguin Form Board. Knox (1914) designed a series of tests for estimating mental defects among immigrants at Ellis Island. Healy and Fernald (1911) produced a series of performance tests, and Pintner and Paterson (1915) adapted a Binet Scale to the deaf. Kelley (1916) devised a construction test, and Dearborn *et al.* (1916) produced a series of form board and construction tests. An elaborate set of form boards was constructed by Ferguson (1920).

There is no clear dividing line between performance tests which use various objects and those which, like the Army Beta, use paper and pencil. Nearly all performance scales include some paper and pencil test of spatial relations.

A. Pintner-Paterson Scale

In 1917 Pintner and Paterson published a standardization of 15 performance tests. Since they were used in the Army testing program and have since been widely applied or adapted, they will be described:

1. The Mare and Foal Test (from Healy). This is a picture of a farmyard from which eleven pieces were cut. The child is presented with the material arranged as shown in Figure 41, 27154, and told, "Put these pieces in the right places as soon as you can." A record is kept of the time in seconds, within a limit of five minutes, and of the number of errors. This same scoring is used in the next ten tests.

2. The Seguin Form Board (Fig. 41, 27159). This is a large

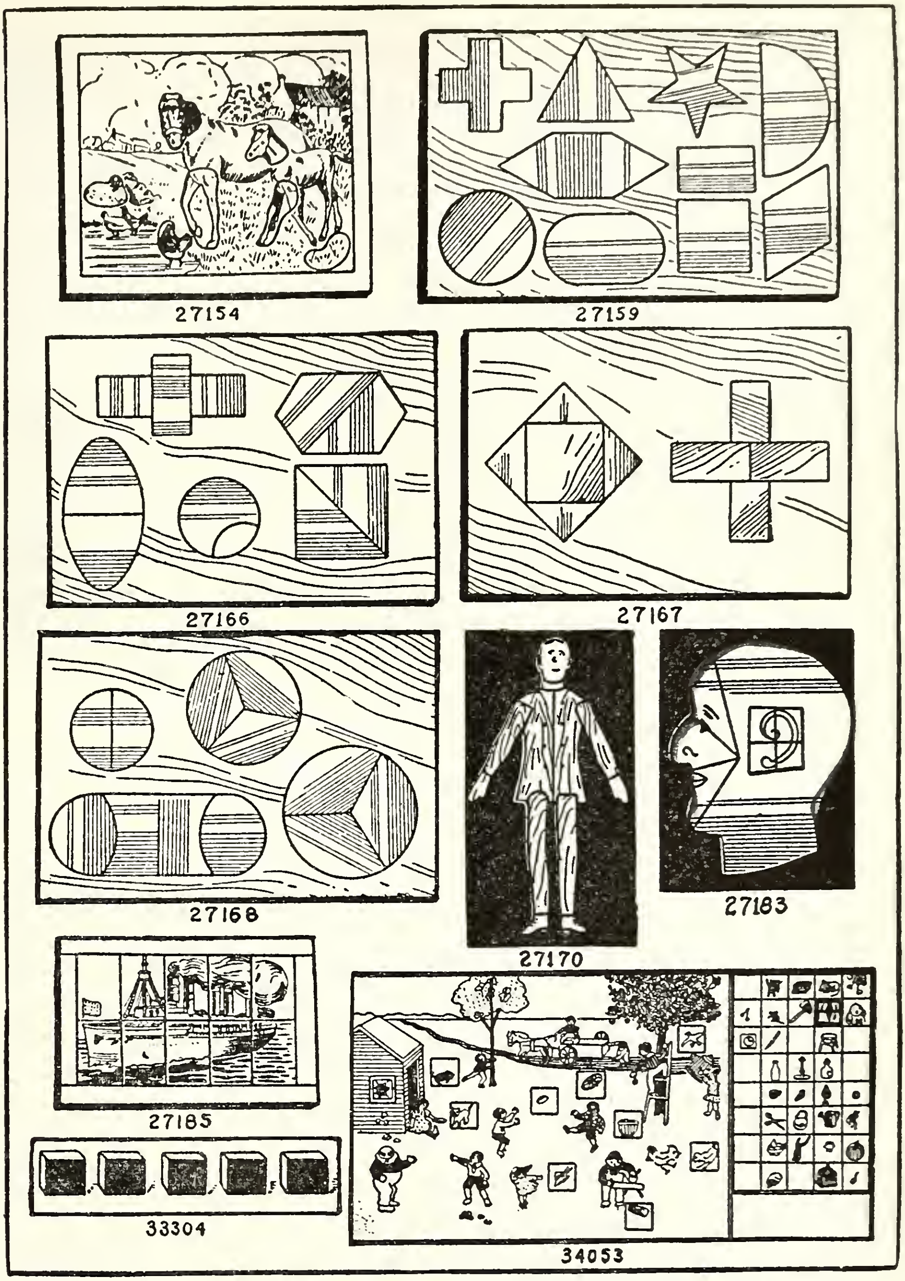

27154
27159
27166
27167
27168
27170
27183
27185
33304
34053

FIGURE 41. **Materials for the Pintner-Paterson Scale**

(Courtesy of the C. H. Stoelting Co., Chicago, Ill.)

board, 20 x 14⅜ inches, from which ten geometrical shapes are cut, Figure 42. The directions and scoring are similar to those in the preceding test. Three trials are given, one right after the other, and results of the best trial are used as the final scores.

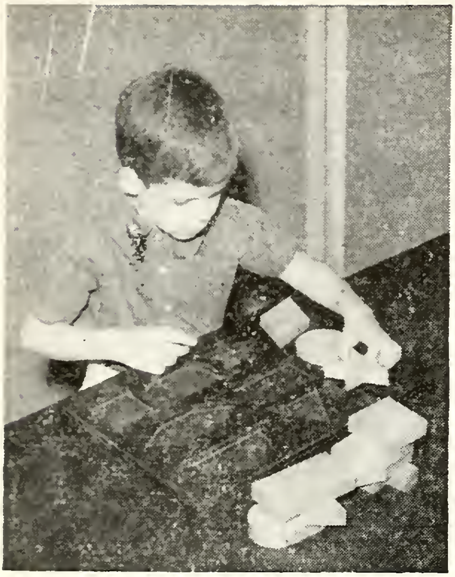


FIGURE 42

Seguin Form Board

Posed by Frank P. Greene

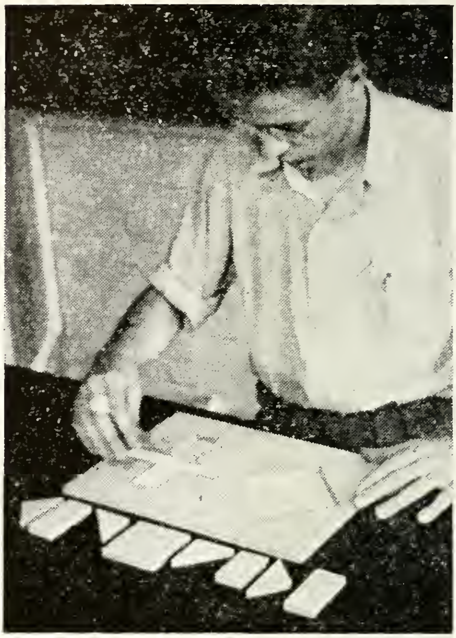


FIGURE 43

Two-Figure Form Board

Posed by Mungo Miller

(Manufactured by the C. H. Stoelting Co., Chicago, Ill.)

3. Five-Figure Board (devised by Paterson), Figure 41. This is similar to the Seguin Board, but more complex in that eleven pieces must be fitted into five holes. One trial was allowed.

4. The Two-Figure Form Board (devised by Pintner), Figure 43. This is similar to the Five-Figure Board, but slightly easier.

5. The Casuist Form Board (from Knox), Figure 44. In this test twelve pieces are to be fitted into four holes. This is considerably harder than the other form boards.

6, 7, 8. These are smaller boards in which four or five pieces are to be fitted into one or two holes.

9. The Manikin Test (Pintner). This consists of a small wooden doll cut in six pieces which are laid out, as in Figure 45. The main difficulty is encountered in making the arms and legs fit exactly, for the places where they fit into the body are not the same shape. This test is for 5-year-olds.

10. Feature Profile Test (Knox), Figure 46. This consists of eight pieces which are to be assembled to form an ear and the profile of a face.

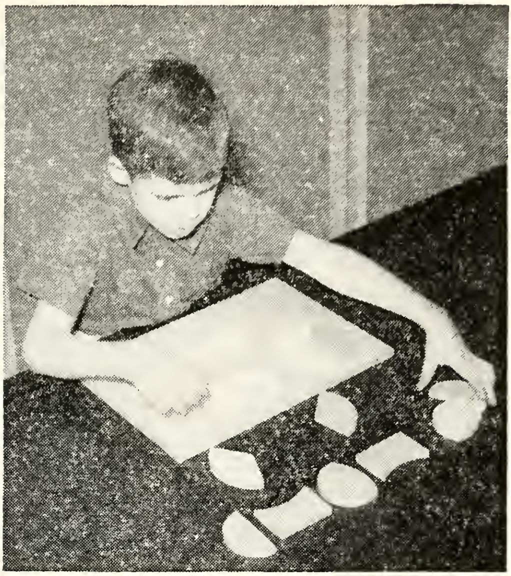


FIGURE 44

Casuist Form Board

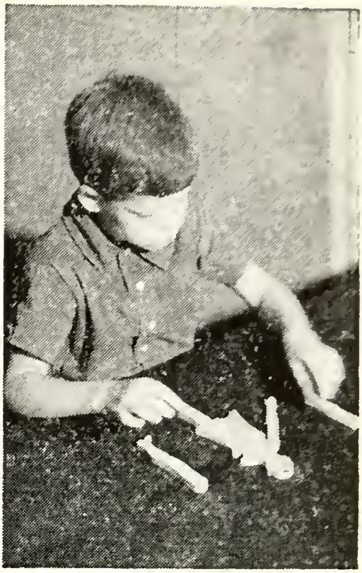


FIGURE 45

The Manikin Test

(Manufactured by the C. H. Stoelting Co., Chicago, Ill.)

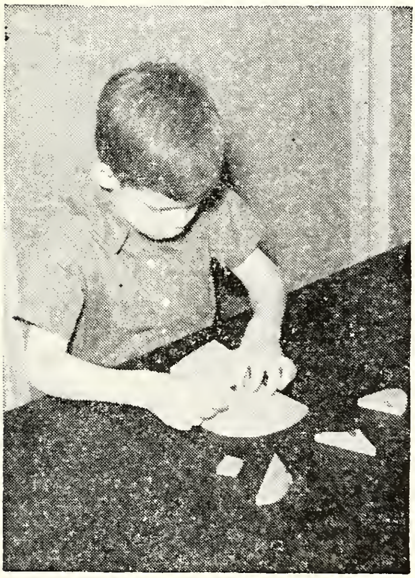


FIGURE 46

The Feature Profile Test

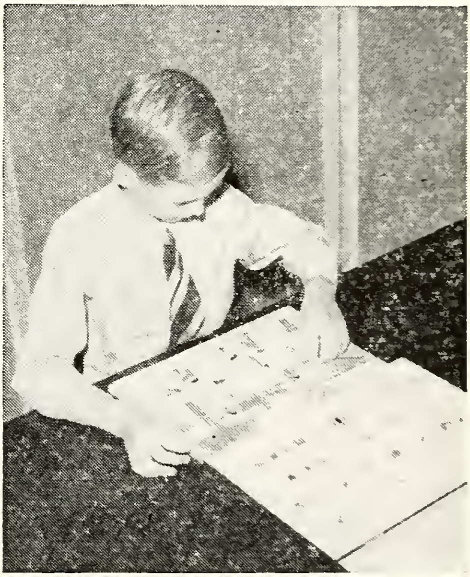


FIGURE 47

The Healy Picture Completion Test, Form I

(Manufactured by the C. H. Stoelting Co., Chicago, Ill.)

11. The Ship Test (after Glueck), Figure 41. This is a picture of a steamship cut into ten rectangular pieces of equal size.

12. The Picture Completion Test (Healy). This consists of a large picture of a rural scene, or of several rural scenes put together. Ten small squares are cut out, as shown in Figure 47. These are to be filled by selecting the most suitable pictures from among forty-eight squares. The score is the number of blanks correctly filled in ten minutes. Actually, five minutes was usually ample time, so the test has been lately standardized with a five-minute time limit.

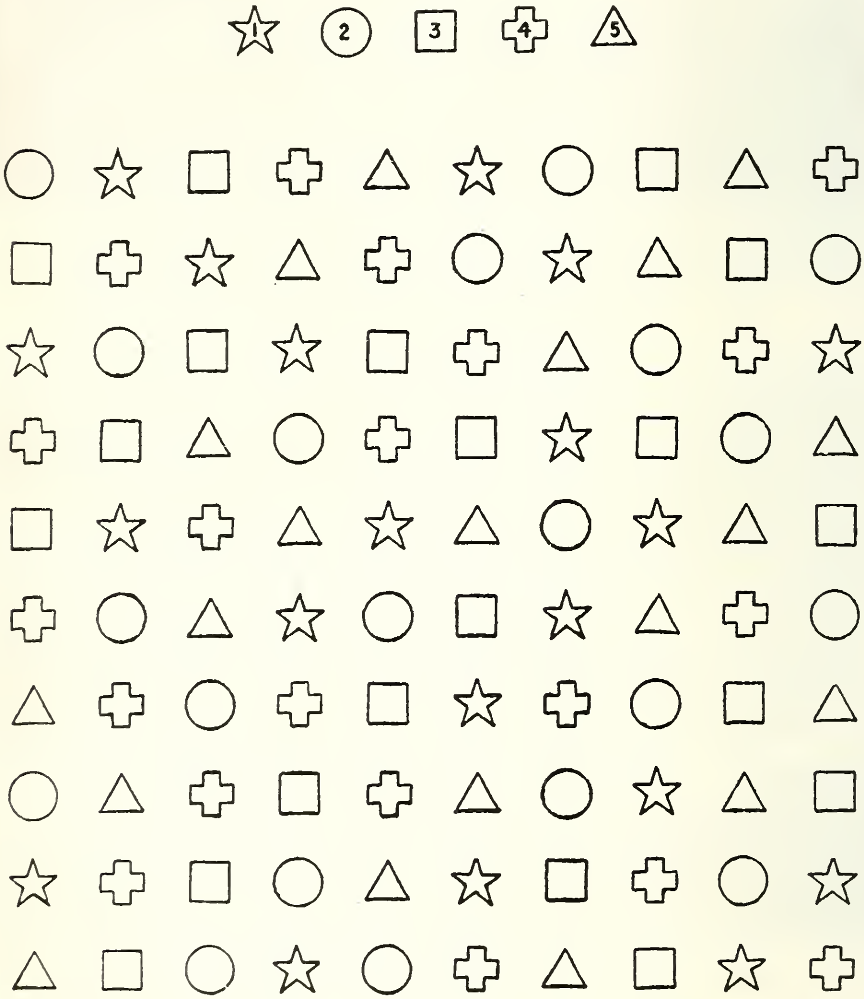


FIGURE 48. **Substitution Test**

(Woodworth and Wells, 1911. By permission of the Editor, *Psychological Monographs*.)

13. The Substitution Test (Woodworth and Wells). This is a paper and pencil test in which rows of five sorts of geometric figures are to be marked with numbers according to the key at the top of the page. The score is the time needed to finish fifty figures (see Figure 48).

14. The Adaptation Board (Goddard). The material consists of a board with four round holes, three of them 6.8 cm. in diameter and the fourth 7 cm. The child is shown how one block fits exactly into the large hole, and then he is asked to "Put it into the right hole" when the board is placed in four different positions. The score is the number of correct trials (see Figure 49).

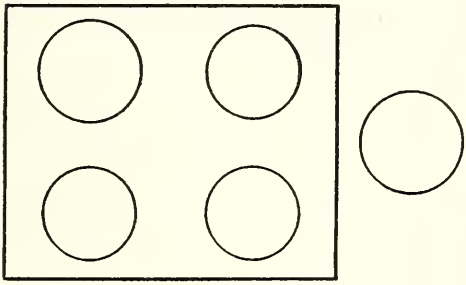


FIGURE 49 (above)

Adaptation Board

(Goddard)

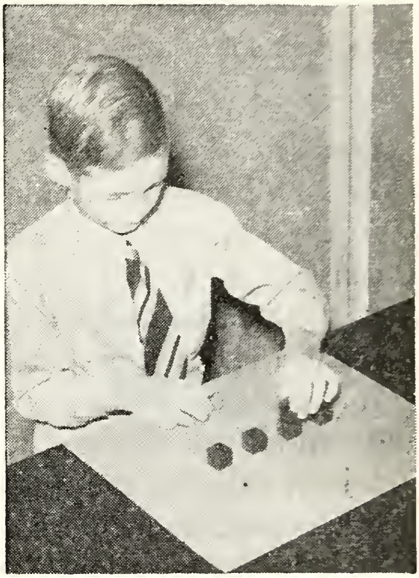


FIGURE 50 (at right)

The Knox Cube Imitation Test

Posed by Donald Johnson

(Manufactured by the C. H. Stoelting Co., Chicago, Ill.)

15. The Cube Imitation Test, Figure 50. The test material consists of five black, one-inch cubes. Four cubes are placed about two inches apart in a row on the table in front of the examinee. With the fifth cube, the examiner taps the other four in a particular order at the rate of one tap a second. Then the examinee is told, "Now you do what I did." The order is simple at first but later becomes more complex. The score is the number of correct trials, which of course reflects the difficult task completed.

The authors prepared a modified age scale and a point scale from their results including time, errors, and total accomplishment. They also published percentiles for each test for each age group and mental-age tables for each test. To secure a single representative mental age, they suggested using a person's median mental age, and this practice has been widely followed. The authors did not offer a systematic analysis of the processes needed for success in these tests, but some aspects seem obvious:

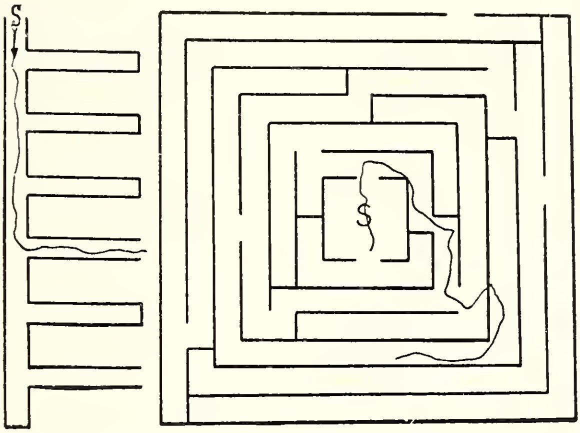
1. Twelve of the tests are speed tests where a few seconds represent the difference between successive age levels.
2. In all tests, manipulation of materials is required, and in the form board precise fitting movements are also required.
3. Perception and comparison of form are essential with the form boards, and ability to interpret pictorial material is appraised by five of the tests, Nos. 1, 9, 10, 11, and 12.
4. In all the tests a systematic procedure or plan is effective in reducing the time needed for success.
5. Immediate memory span is stressed in the Cube Test and the Substitution Test.

A short form of this scale, including Tests 1, 2, 3, 4, 5, 9, 10, and 11, has recently been issued by Pintner and Hildreth (1937).

B. Arthur Performance Scale

In 1930 Arthur published a restandardization of ten of the tests in the Pintner-Paterson scale, Nos. 1, 2, 3, 5, 7, 9, 10, 11, 12, and 15, based on scores of approximately 1,100 school children from 6 to 16 years of age. She also restandardized the Porteus Maze Tests and the Kohs Block Design Scale.

The Porteus Maze Test (1924) requires one to draw with a pencil through printed mazes, see Figure 51 A and B. Eleven mazes have been prepared for age levels from 5 years to superior adult. The mazes are called roadways, and one's pencil line must not cross any printed lines nor go into any closed roads. Failure to observe these instructions leads to the immediate removal of the printed maze and a second trial with a duplicate form. Only two trials are allowed in nine of the mazes, but in the 12- and 14-year levels four trials are given. Success on the later trials counts less than success on the first. The total credit is given in years and months of mental age. No time limits are set. Speed is not a factor in this score, but caution is probably a large factor. This test was adapted for group testing in the Army Beta, page 311, using a time limit which made speed of working very important.



Year V - 2 Trials

Year XIV - 4 Trials

FIGURE 51A

Porteus Mazes

(Manufactured by the C. H. Stoelting Co., Chicago, Ill.)

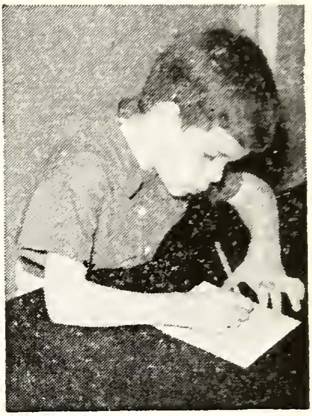


FIGURE 51B

Porteus Maze Test

Posed by Frank P. Greene

The Kohs Block Design Scale (1927) requires the examinee to duplicate with colored cubes the designs on 17 printed cards. All the cubes are the same, each having four sides colored red, white, blue, and yellow, respectively; and the other two sides being divided diagonally between blue and yellow, and red and white, respectively. After a simple demonstration, the cards are presented with the required number of blocks. The simplest designs use only four blocks and the most complex, sixteen blocks. Time is limited to short working periods for each design, and the scores are points assigned on the basis of speed of completing a design.

Arthur's scoring ignored errors or number of moves, but allowed credit for either speed of performance or, in the untimed tests, achievement. Weighting the scores for errors was not found to have much effect upon one's position in a group of persons. Arthur published age equivalents for the scores of each test. She also furnished point scales which assigned credit on the basis of the power of a test to discriminate between adjacent age groups. By this procedure the relative weight allowed for a particular test may vary from year to year. She advised the summation of points to give a total score for the test. This total may be converted to a performance age (P.A.) by reference to a table.

C. Other Scales

A number of the other nonverbal scales designed to measure general ability should be mentioned. Pintner (1919) and, more recently, Brown (1936) have developed paper and pencil tests which follow somewhat the Army Beta style and which extend the range to either lower or higher levels of ability. Dodd (1926) developed an elaborate nonverbal test which consisted of arranging pictures in time sequences, selection by matching pictorial relations, and a modified maze. Greene (1931) developed and standardized two nonverbal tests, Figure 52, which were intended to measure observation and comparison with simple materials. One test requires the selection and marking of a unique feature in a row of small faces. The other test is a series of mazes which range from easy to very difficult. The scores on these two tests are the number of items correct in a limited time, and the number of errors. No penalty for errors was imposed, for it was thought that errors were more determined

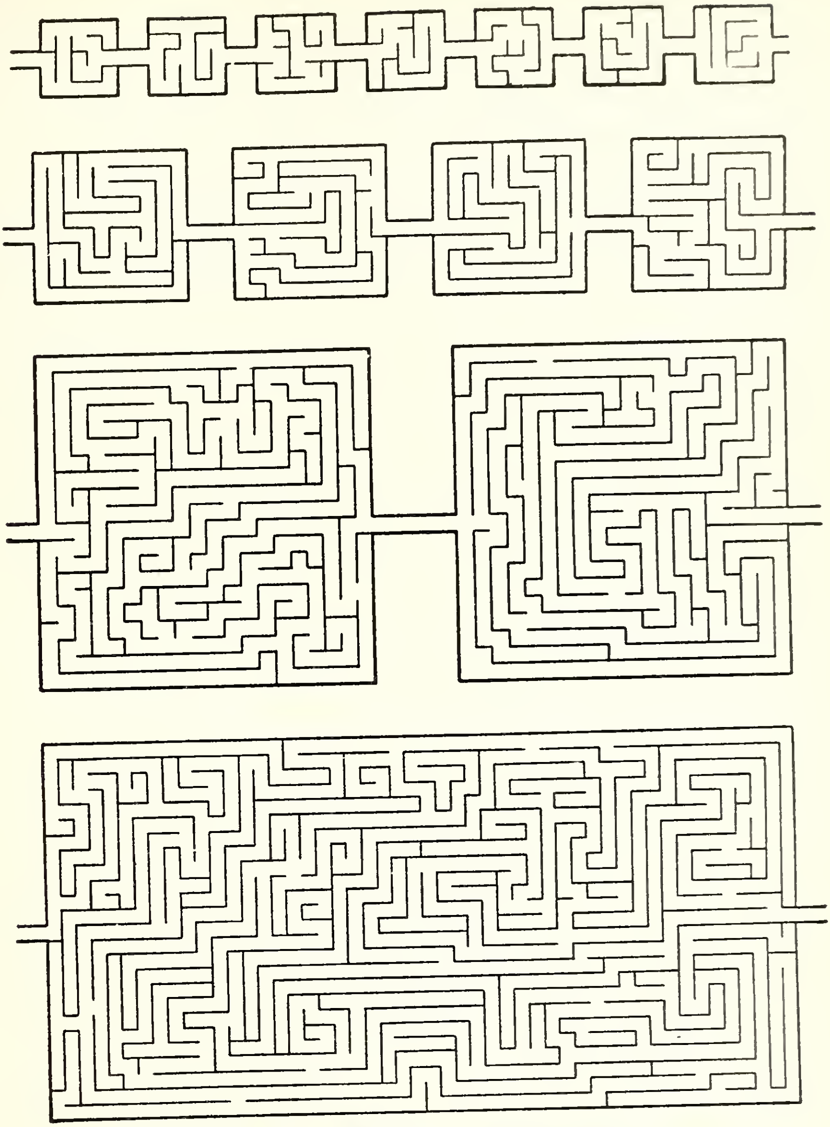


FIGURE 52. **Michigan Nonverbal Series**

Samples from separate pages. Each level of complexity was scored separately.

(Greene, 1931)

by temperament than by intellectual ability. For every age group, Greene reported nearly zero correlations between achievement and number of errors. Since errors seem to result from carelessness or from going too fast, it seemed sensible to give them separate scores rather than to combine them with achievement scores.

III. MECHANICAL ABILITY TESTS

The term *mechanical ability,* like the term *general intelligence,* has numerous and somewhat vague meanings. Historically, mechanical ability has been defined as that which enables a person to work with tools and machinery. Tests of mechanical ability may be grouped into two classes: those which require information or reasoning about tools or machinery, and those which require in addition, the manipulation of machines. The former are frequently paper and pencil group tests, and the latter are usually individual assembly or stripping tests.

A. Paper and Pencil Tests

Mechanical tests of this sort are very widely used in appraising achievement in manual training, and in predicting success in future training or employment. An elaborate set of trade questions were constructed for use in the United States Army in 1917 and proved to be valuable for classification of newly recruited soldiers. Tests of this sort have been copied and expanded both in private industry and in public personnel administration. A few verbal and pictorial samples will be described.

Verbal tests of mechanical information are illustrated by the multiple-choice items concerning woodworking shown in Table 66. These are taken from an elaborate study of mechanical ability by Paterson *et al.* (1930) , which included appraisals of knowledge in each of the following: woodworking, printing, mechanical drawing, sheet metal working, and electricity. A large number and variety of examinations of this sort are prepared by teachers of mechanical subjects in high school and college and by civil service examiners. Norms have also been published for a rather easy sixty-item

TABLE 66

Manual Training Information Tests

Minneapolis Public Schools

Test A. Woodwork

Devised by Manual Training Department Committee on Objectifying Grades on Woodwork, assisted by the Mechanical Abilities Research Staff, University of Minnesota.

Name..............................School..............................

Date of birth..Grade..................
day month year

Underscore shop courses you have taken: Sheet Metal, Mechanical Drawing, Electricity, Woodwork, Printing.

This is a test to see how much you know about Woodwork.

Here is a sample question already worked out. Notice how it is done.

St. Paul is the capital of 1. Ohio 2. Vermont 3. Minnesota.........(3)

The right answer is Minnesota, so Minnesota is underlined. Notice also that the word Minnesota is No. 3, so 3 is written in the parenthesis at the right hand side of the page. You are not expected to answer all the questions below. Answer just as many of them as you can.

Underline only one word in each case and be sure to put the number of that word in the parenthesis at the right-hand side of the page.

Stop. Wait for signal before beginning.
Test begins here.

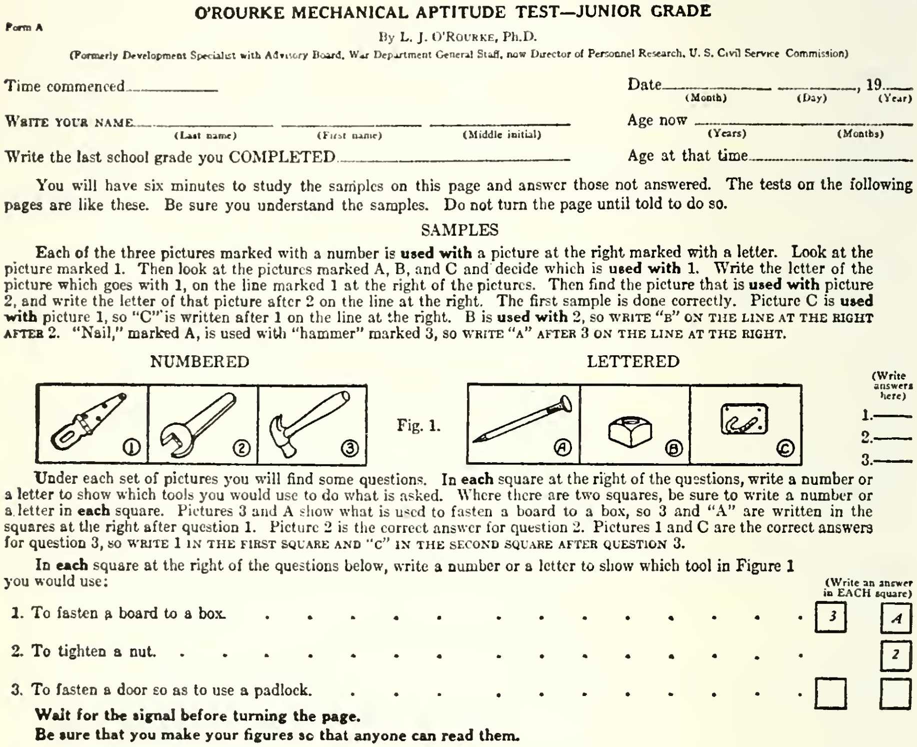
1. For holding pieces of wood together while gluing, use
   1. clamps. 2. screws. 3. boards. 4. nails................( )
2. Holes are bored with
   1. a brace and bit. 2. an awl. 3. a planer. 4. a chisel...( )
3. A jack plane is made of
   1. steel. 2. copper. 3. celluloid. 4. glucose.............( )
4. A part of a brace is the
   1. chuck. 2. point. 3. break. 4. knife..................( )
5. The marking gauge is used in
   1. laying out widths. 2. cutting small boards. 3. making joints tight. 4. boring holes...................................( )
6. When sanding flat surfaces, the sandpaper should be backed with a
   1. flat block. 2. wooden plane. 3. cylinder. 4. piece of pumice stone..( )

(and 132 more items)

(From Paterson *et al.*, 1930, p. 150. By permission of the University of Minnesota Press.)

test of knowledge of machinery, woodworking, and painting by O'Rourke (1936), and for tests of *general mechanical ability* by Baker and Crockett (1928) and by Baker and Voelker (1937).

Pictorial tests of mechanical information are illustrated by Figures 53 and 7, and by Table 64. In Figure 53, pictures of small objects are to be marked to show which are used together. In Figure 7, page 30, the parts of a lathe are to be labeled and their functions explained. In Table 64, section 8A, page 333, the task is to indicate the movement of a part of a machine which would occur upon a movement of some other part.



Form A

O'ROURKE MECHANICAL APTITUDE TEST—JUNIOR GRADE

By L. J. O'Rourke, Ph.D.

(Formerly Development Specialist with Advisory Board, War Department General Staff, now Director of Personnel Research, U. S. Civil Service Commission)

Time commenced ____________

Date ______ (Month) ______ (Day), 19___ (Year)

Write your name ______ (Last name) ______ (First name) ______ (Middle initial)

Age now ______ (Years) ______ (Months)

Write the last school grade you COMPLETED ____________

Age at that time ____________

You will have six minutes to study the samples on this page and answer those not answered. The tests on the following pages are like these. Be sure you understand the samples. Do not turn the page until told to do so.

SAMPLES

Each of the three pictures marked with a number is **used with** a picture at the right marked with a letter. Look at the picture marked 1. Then look at the pictures marked A, B, and C and decide which is **used with** 1. Write the letter of the picture which goes with 1, on the line marked 1 at the right of the pictures. Then find the picture that is **used with** picture 2, and write the letter of that picture after 2 on the line at the right. The first sample is done correctly. Picture C is **used with** picture 1, so "C" is written after 1 on the line at the right. B is **used with** 2, so WRITE "B" ON THE LINE AT THE RIGHT AFTER 2. "Nail," marked A, is used with "hammer" marked 3, so WRITE "A" AFTER 3 ON THE LINE AT THE RIGHT.

(Write answers here)

1. ______
2. ______
3. ______

Under each set of pictures you will find some questions. **In each** square at the right of the questions, write a number or a letter to show which tools you would use to do what is asked. Where there are two squares, be sure to write a number or a letter in **each** square. Pictures 3 and A show what is used to fasten a board to a box, so 3 and "A" are written in the squares at the right after question 1. Picture 2 is the correct answer for question 2. Pictures 1 and C are the correct answers for question 3, so WRITE 1 IN THE FIRST SQUARE AND "C" IN THE SECOND SQUARE AFTER QUESTION 3.

In **each** square at the right of the questions below, write a number or a letter to show which tool in Figure 1 you would use:

(Write an answer in EACH square)

1. To fasten a board to a box. . . . . 3 A
2. To tighten a nut. 2
3. To fasten a door so as to use a padlock. . . . . ☐ ☐

Wait for the signal before turning the page.

Be sure that you make your figures so that anyone can read them.

FIGURE 53. **A Mechanical Perception and Information Test**

(By permission of the Psychological Institute, Washington, D. C.)

Pictorial tests of ability to observe and compare geometric shapes or other spatial patterns are shown in Figures 54 and

READ THE FOLLOWING DIRECTIONS VERY CAREFULLY WHILE THE EXAMINER READS THEM ALOUD

Look at the problems on the right side of this page. You will notice that there are eight of them, numbered from 1 to 8. Notice that the problems go DOWN the page.

First look at Problem 1. There are two parts in the upper left-hand corner. Now look at the five figures labelled A, B, C, D, E. You are to decide which figure shows how these parts can fit together. Let us first look at Figure A. You will notice that Figure A does not look like the parts in the upper left-hand would look when fitted together. Neither do Figures B, C, or D. Figure E does look like the parts in the upper left-hand corner would look when fitted together, so E is PRINTED in the square above $\boxed{1}$ at the top of the page.

Now look at Problem 2. Decide which figure is the correct answer. As you will notice, Figure A is the correct answer, so A is printed in the square above $\boxed{2}$ at the top of the page.

The answer to Problem 3 is B, so B is printed in the square above $\boxed{3}$ at the top of the page.

In Problem 4, D is the correct answer, so D is printed in the square above $\boxed{4}$ at the top of the page.

Now do Problems 5, 6, 7, and 8.

PRINT the letter of the correct answer in the square above the number of the example at the top of the page.

DO THESE PROBLEMS NOW.

If your answers are not the same as those which the examiner reads to you, RAISE YOUR HAND.

DO NOT OPEN THE BOOKLET UNTIL YOU ARE TOLD TO DO SO.

Some of the problems on the inside of this booklet are more difficult than those which you have already done, but the idea is exactly the same. In each problem you are to decide which figure shows the parts correctly fitted together. **Sometimes the parts have to be turned around, and sometimes they have to be turned over in order to make them fit.** In the square above $\boxed{1}$ write the correct answer to Problem 1, in the square above $\boxed{2}$ write the correct answer to Problem 2, and so on with the rest of the test. Start with Problem 1, and go DOWN the page. After you have finished one column, go right on with the next. **Be careful not to go so fast that you make mistakes. Do not spend too much time on any one problem.**

PRINT WITH CAPITAL LETTERS ONLY.

MAKE THEM SO THAT ANYONE CAN READ THEM.

DO NOT OPEN THE BOOKLET BEFORE YOU ARE TOLD TO DO SO.

YOU WILL HAVE EXACTLY 20 MINUTES TO DO THE WHOLE TEST.

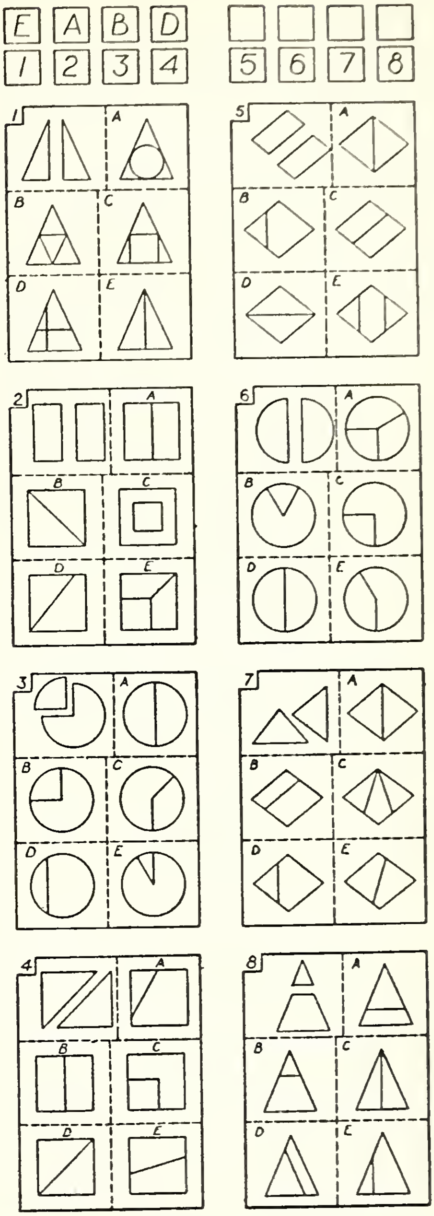


FIGURE 54. **The Revised Minnesota Paper Form Board**

(Courtesy of Likert and Quasha, 1934.)

55. Tests of this sort are often included in both performance tests and tests of mechanical aptitude. Figure 54 shows a recent revision of a paper form board, a simpler form of which is shown in the United States Army Beta, Test 7,

page 317. The revision is more difficult than the original and also easier to score, because it eliminates the interpretation of a line drawing. Figure 55 is a test of counting pictures of piles of blocks. It involves enumeration as well as visual imagination of the unseen portion of the pile.

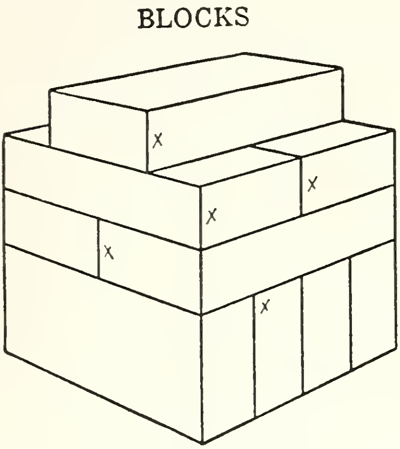


FIGURE 55

Counting Blocks

The task is to indicate the number of blocks which touch each block marked X.

(MacQuarrie, 1925. By permission of the California Test Bureau.)

B. Assembly Tests

Tests of assembling and stripping mechanical objects involve manual operations upon a standard piece of apparatus. There does not appear to be a sharp boundary line between the tests of dexterity which are described below and the tests of mechanical assembly which are included in this section. Both involve speed and accuracy of hand and eye coordination, but the latter also demand information and skill in problem solving. In dexterity tests the score is principally determined by motor co-ordination, but in the assembly tests motor co-ordination may not be an important factor in success.

One of the earliest tests of mechanical assembly is the Puzzle Box of Healy and Fernald (1911). The glass-framed lid of a small wooden box may be opened by releasing, with a button hook, strings which were secured over pegs. The order in which the strings must be released can be determined by visual examination of the box, inside and out. A similar puzzle box was designed by Freeman (1916) in which the eleven levers had to be moved in a particular sequence in order to open the box (see Figure 56). Both of these tests showed large differences between sexes, and five-minute periods were found to be too short for average adults.

Another test of this sort is O'Connor's (1928) Wiggly-Block Test. Nine similar pieces of wood, cut with some wavy edges, are to be fitted together to form a solid block 9 × 9 × 12 inches. The score is the time needed to finish the task.

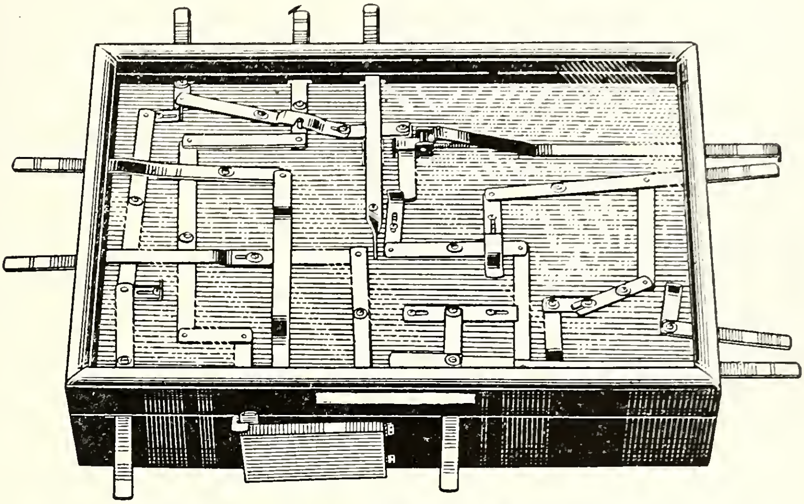


FIGURE 56. **Freeman's Puzzle Box**

(Courtesy of the C. H. Stoelting Co., Chicago, Ill.)

None of these three tests have shown high enough reliabilities to be considered good measuring techniques. The low retest reliabilities probably reflect the fact that some persons take a long time to puzzle out the solution on the first trial but have a good memory of the solution in the second trial, and that others may enjoy some chance successes on the first trial but fail to remember the solution later.

The Minnesota Spatial Relations Test, Figure 57, uses an elaborate set of four form boards, each of which has about sixty pieces varying in size or shape. This is an extension of a similar test devised by Link (1919). Patterson *et al.* (1930) found that Link's test was too short to give suitable reliabilities; so they made the test about eight times as long and achieved a retest reliability of .84 on a sample of 217 boys. The score is the number of seconds needed to complete the assembly, and the task, at least for adults, is largely one of visual comparison of form and methodical work. Dexterity was found to be relatively unimportant by Harrell (1939).

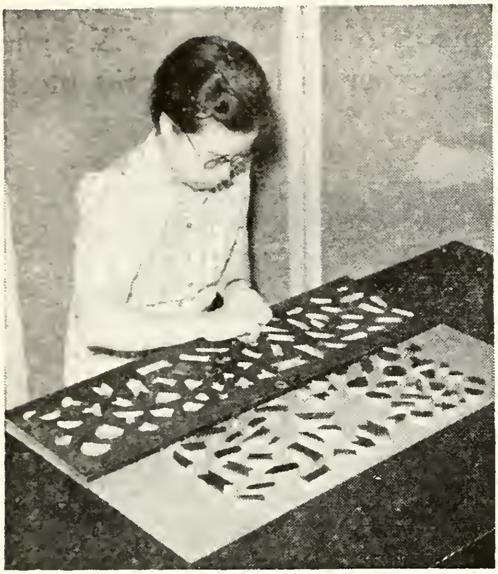


FIGURE 57

The Minnesota Spatial Relations Test

Posed by Dr. Bing Chung Ling

(Manufactured by the Educational Test Bureau, Minneapolis, Minn.)

Another type of mechanical manipulation test requires the assembly or stripping of common hardware. Stenquist (1923) found norms for the assembly of ten small objects which could be purchased at local stores. The objects which he used were:

1. A clothes pin with a wire spring
2. A Hunt paper clip
3. Rubber hose shutoff
4. Chain with split links
5. Bicycle bell
6. Wire bottle stopper
7. Push button
8. Small door lock
9. A cupboard latch
10. Mouse trap

The objects, taken apart, are presented to the examinee one at a time in order of difficulty. A small screw driver is available. The score is the number of correct assembly operations completed in 30 minutes.

Paterson *et al.* (1930) believed that the odd-even reliability of the Stenquist Assembly Test, which was found to be .72 among 217 seventh and eighth grade boys, was not high enough for careful appraisals. Therefore they modified and enlarged the material to include thirty-six items, Figure 6, page 28. Success in each of these, which was scored on a ten-point scale, was found to correlate .28 or better with total scores. The highest correlations with totals were found to be in the assembly of the bicycle bell, .75; the clothes pin, .68; spark plug, .66; and chain, .64. The odd-even correlation of the total test (using the Spearman-Brown prediction formula) was .94. Retest reliabilities, which would doubtless be somewhat lower, have not yet come to hand.

An assembly test for girls was constructed by Toops (1923) using the same scheme as the Stenquist Assembly Test but with different materials. Eleven items were included:

1. Stringing beads
2. Inserting tape
3. Rosette
4. Cross-stitch
5. Key ring
6. Clips
7. Tape sewing
8. Trunk tag
9. Card wrapping
10. Booklet
11. Cut-out of trimming

For each item a finished sample was presented and also the materials necessary for the construction of another sample. Scores were accumulated from credits assigned to various achievements during a 45-minute period. Metcalfe and Burr (1937) found that it was possible, by the use of a shorter battery of seven of these tests, to secure a score which correlated .93 with the complete battery among 1,300 girls in an industrial group. They therefore eliminated the four tests which were most troublesome to administer and score. The seven remaining were:

2. and 3. Simple cloth assembling
4. and 7. Facility in the use of a needle
6. and 9. Difficult assembling, using clips, paper, and string
11. Use of scissors

These same authors reported in 1938 that the mean correlation between MA's and the girls' mechanical assembly test scores (GMA's) was .40, although the mean correlation between chronological age (CA's) and GMA's was nearly zero. These results were related to the fact that the older girls in the sample were of lower general mentality. In order to provide meaningful norms, the authors used MA groups rather than CA groups. The mean score ranged from 22 for girls with a mental age of 8 years, to 51 at MA 16 years. The standard deviation of the younger group was 13.8, and of the older group, 10.5. This reduction in SD is probably a reflection of both a small sampling of the population and a scale which was too easy to discriminate among the highest third of the higher groups. The test has, however, proved to be effective in predicting success among a number of factory jobs which require routine piece part assembly and simple machine operation.

None of the assembly tests thus far described have included appraisals of methods of work, although several writers have pointed out that large individual differences exist. An illustration of an evaluation of procedures is seen in the work of Farmer (1930), who watched 259 boys perform a cube construction test and made full notes of the method employed. From these observations there appeared to be five common and industrially important modes of behavior which he called (p. 46):

> *The controlled type.* The chief characteristic of this type was their complete mastery of the situation. They did not appear to think ahead, nor were they worried by the test in any way. They did it without any concern, hurry, or manifestation of interest.
>
> *Thinkers.* This category was confined to what appeared to be abstract thinkers. They seem to think ahead and have a definite plan in their mind before doing the test.
>
> *Good workers.* This category consisted of the conscientious plodding type. However much they were baffled by the problem presented they continued to work at it. Sometimes they broke down all they had done and started again. They were often worried by the test, but they never thought of giving up and they learnt by experience.

The inept. This category was composed of those who had no idea whatever how to do the test. (In my original notes I called them "fools," and that is really the best description, for their methods seemed to be utterly irrational.)

Miscellaneous. Into this category I put those about whom I was uncertain, or who failed to impress me with any outstanding characteristic. The results obtained from the first group of subjects showed that it was better to have such a category rather than force every one into one of the positive categories.

After three years' training in a school for industrial apprentices, the industrial proficiency of boys was measured by a practical examination. These proficiency grades were compared with earlier records of behavior on the cube test. The first three types of performers – the *controlled type,* the *thinkers,* and the *good workers* – were found to have succeeded better on the average than the last two types – the *inept* and the *miscellaneous.* The biserial correlation of proficiency with mode of behavior was .23; with scores on the cube test, .228; and with a psychological test, .37. Farmer concluded that the method contributed slightly to the prediction of success, but that this preliminary work should be refined to make more accurate predictions.

A number of assembly tests have been described by Russian, German, French, and British industrial psychologists which cannot be evaluated here. In this country there are also a number of complex form boards and assembly tests which have been used locally, but for which data are not yet available.

IV. MOTOR CO–ORDINATION TESTS

The field of motor co-ordination has received more attention in laboratory studies than in the standard testing techniques. This situation is probably due to the fact that thorough motor testing employs rather extensive mechanical equipment and demands individual administration. A few fairly well-standardized procedures using simple materials have been developed. These will be discussed under four headings: (*a*) Reaction Time, (*b*) Agility and Strength, (*c*) Dexterity, and (*d*) Steadiness Tests.

A. Reaction Time

Reaction times have been extensively studied for both total and partial responses. A total reaction time is the period between the application of a stimulus, such as the sound of a bell, and some muscular response, such as the release of a telegraphic key. A partial reaction time is a period between the application of a stimulus and a change in some portion of the total reaction path.

Numerous studies of total reactions of the simplest sort have been extensively discussed by Woodworth (1938). The evidence is conclusive that for adult laboratory subjects:

1. Reaction times vary with sense organ stimulated. Approximate mean reaction times in seconds, as given by moving a finger as quickly as possible after a moderate stimulus, were for:

Touch on the hand	.120	Pain	1.000+
Touch on the forehead	.130	Smell	1.000+
Sound	.130	Salt taste	.307
Light	.180	Sweet taste	.496
Cold	.150	Acid taste	.536
Warmth	.180	Bitter taste	1.082

2. Reaction times vary with the intensity of the stimulus; in general, the more intense the stimulus, the quicker the response.

3. Reaction times vary with the nervous connections. Thus, the response of the right hand to stimulation of the right hand is faster than to the stimulation of the left hand or of either foot.

4. The reaction time varies with the reacting movement. A well-practiced movement is quicker than an unfamiliar movement. The force used to close a switch is much greater than necessary at first, but later in practice a smaller well-directed movement is used.

5. Ready signals are very important. Regular signals about one-half second before the stimulus result in shorter reaction times than longer, irregular, or shorter signals.

6. Discrimination reactions, such as pressing a key if a sound

comes from the left but not if it comes from the right, are much slower than simple reactions.

7. Associative reaction times, such as calling out the first word that is suggested by a stimulus, vary greatly. The more familiar associations are nearly as fast as simple response time to visual stimuli, but the less familiar or the emotionally toned responses are much longer.

Van Essen (1935) reported a number of interesting observations in connection with studies of auto drivers. Manual reaction times to auditory and visual stimuli varied with the type of stimulation, thus:

1. A short stimulation gave slightly shorter reaction times than the disappearance of a continuous stimulus, and a much shorter reaction time than the appearance of a continuous stimulus.
2. The reaction times became shorter when subjects' reactions affected the stimulus, as, for instance, when lifting one's foot from a pedal extinguished a light, than when the reaction did not affect the stimulus.
3. Reactions to a more distant red light (24 meters) were faster than to nearer lights (12 meters or 2 meters) when intensity and size of retinal stimulation were held constant.
4. Repeated reaction times did not fall into a normal curve of distribution for each individual, but into several patterns which depended upon qualitative differences in reaction patterns.

One practical standardization of reaction times is reported in Chapter 21 for a study of senescence by Miles (1931). Another study, showing growth and practice norms, is that of Jones (1937). He instructed his subjects:

> This is a test to see how fast you can move your hand. You place your hand on this board and hold it down: in a moment a light will appear in the red bulb. This is a signal that in 1, 2, or 3 seconds a buzzer in the clock will sound. When you hear that buzzer, you must lift your fingers off just as fast as you can (demonstrating). At the sound of the buzzer, a little electric clock starts; the clock stops when you lift your fingers, and by reading the dial we can tell how quick you are.

After the preliminary trials, 15 trials were taken with each hand. Some of the results were:

1. Odd-even reliability correlations for 15 trials were approximately .87 for boys and .89 for girls.
2. No marked differences were found between mean scores for right and left hands. Correlations between right and left hand responses ranged from .80 to .86.
3. Retest correlations on groups of 90 children after a year's interval were from .60 to .72 and after 3 years, from .55 to .57.
4. Small practice effects occurred in 11-year-old children on the second day but not thereafter during 4 days of practice.
5. Males were slightly superior to females, particularly for the left hand and at the earlier ages.
6. A marked warming-up effect was noticed during the first five trials with the right hand and during all 15 trials of the left hand.
7. Age norms for right-hand reaction times were:

		Mean
4.5 years	(37 cases)	398.3
7.3 years	(34 cases)	250.8
10.9 years	(76 cases)	186.7
14.7 years	(50 cases)	162.5
19.5 years	(40 college students)	156.7

8. There was some evidence that adaptation was present in groups where the test was repeated four times at yearly intervals. These groups surpassed their controls.
9. Motivation was found to be very important. The subjects were allowed to see their own reaction times on the face of a chronoscope and were urged to break their own records.

Studies of total reaction times have convinced investigators that a response was a complicated affair which should be analyzed into its constituent parts. An early analysis was made by Exner (1868), who described seven segments of a reaction:

1. Time needed for excitation of sense organ
2. Conduction through a sensory nerve
3. Conduction from lower spinal cord to lower brain centers
4. Conduction from sensory to motor brain centers
5. Conduction from motor center to lower cord
6. Conduction through motor nerve
7. Muscular movement

These or similar segments have been studied by means of small electrical conductors placed in or near various parts of the nervous pathway. One variety of this technique measures the time between a particular stimulus and the changes in the resistance of the skin on some part of the body. This reaction is often called a psychogalvanic reflex. An excellent review of research dealing with the galvanic skin responses is given by Woodworth (1938). Another variation of the technique records small changes in electrical resistance which occur in brain tissue or in the scalp. Photographic records of such changes are called electro-encephalographs. This fascinating field, which is being rapidly developed, is reviewed by Gibbs (1937), Berger (1937), Davis (1938), and others.

B. Agility and Strength

The tests for infants described in Chapter 8 were found to contain large sections devoted to appraisals of posture, hand and eye co-ordination, and locomotion. Tests of similar patterns of behavior have been developed for older children and adults, and the norms in some cases are fairly adequate.

A series of motor tests for children from 2 to 6 years of age were tentatively standardized on 98 children by McCaskill and Wellman (1938). They describe the tests as follows (p. 141):

> It was the idea from the beginning to keep each activity in the test situation as simple as possible so that it could be reproduced easily. For the ball throwing and ball bouncing a "location field" was devised for determining the distance and direction of the child's throw or bounce. This field was made of heavy brown paper 8 feet wide and 17 feet long, with a 4 inch strip of wall board at each end to anchor it. The width was marked off into zones. Zone 1 was 2 feet in width and extended down the center of the field, with zones 2, 3, and 4, each 1 foot wide, to the right and left of zone 1. The length was marked off into distances; distance 1 was 3 feet in length and distances 2, 3, 4, 5, 6, 7, and 8 each 2 feet in length. For the throw or bounce the child

stood at the edge of and on the center line of the field and threw or bounced the ball to the experimenter at the opposite end of the field. The results were recorded according to zone and distance and the child's use of one or both hands.

In the catching series the ball was thrown to the child at a level with his chest each time, as nearly as was possible. The method used in the attempt to catch the ball, the successes, and whether or not the child used defense movements when the ball was tossed to him were all recorded. Two balls were used throughout, one 16¼ inches in circumference, the other 9½ inches. Three trials were given for each performance with each ball.

To determine the child's ability to maintain equilibrium and balance, a walking path and a circle were used. A demonstration was given in each case. If the child failed to respond, he was asked to follow the experimenter as she walked on the path or circle. The path, 10 feet long and 1 inch wide, was drawn on a large piece of brown paper and colored red. The circle was cut from wall board and was 4 feet in diameter. A strip (1 inch wide) was drawn around its outer border and colored red like the path. The number of times the child stepped off the path and circle on each trial was recorded. Three trials were given on each. If the child set one foot off the line for balance but did not take a step, he was not penalized.

Four heights were used for the jumping – boxes 8, 12, 18, and 28 inches in height. The child was given three trials at each height and was checked according to the method he employed in the jump.

The child was asked to hop on one and on both feet and a demonstration of each was given him. He was given three trials at each. The stages were recorded in number of steps. Four items were listed under skipping: walking, shuffle, skipping on one foot, and skipping on alternating feet. The experimenter skipped and asked the child to follow her. The shuffle has a rhythmic quality – the same foot is always forward at each advance. It is definitely in advance of walking and is not the same high step the children employ in galloping.

Both ascending and descending steps were tested on short and long flights. The short flight had 4 steps 7 inches in height with an 11 inch tread. The long flight had 11 steps of the same height and tread. The rail on both flights was 29 inches in height. The steps on which the kindergarten children were tested were slightly different. The short flight had 4 steps 7 inches in height with a tread of 10 inches and the long flight 12 steps 6¼ inches in height with a tread of 11¾ inches. The hand rail was 34 inches above the steps. The child had three trials on ascending and descending both flights of steps.

For ladder climbing two ladders were used. One had twelve rungs 6 inches apart, the other six rungs 12 inches apart. The ladders were placed at approximately a 45 degree angle each time. This placement was kept constant by placing the same ladder rung against the support each time.

In scaling these tests, months of motor age were assigned to each observed stage of development, or item, from the results of testing the standardization group. Age equivalents were calculated using Thurstone's (1925) method (see Chapter 22), which shows the age at which 50 per cent of the group passed the item, and 50 per cent failed to pass. Table 67 shows the motor ages for each item, which, incidentally, correspond very closely to values found by Bayley (1935) for hopping, walking, and stair climbing among 2- and 3-year-olds. Table 67 also shows a point scale used for individual scores. Separate scores were provided for four kinds of activities by summing the points made on groups of items called (1) *Steps and Ladders,* (2) *Ball Activities,* (3) *Jumping,* and (4) *Hopping, Skipping, and Walking.* Total score norms were also published. The mean correlations between these types of motor skills for all ages combined were approximately .69 for boys and .75 for girls. The true correlations between these skills would doubtless be much lower if age were held constant. The lowest correlations were between *Ball Activities* and *Steps and Ladders,* and the highest between *Steps and Ladders* and *Hopping.*

Forty-six youngsters were retested within one week. The total score retest reliability was .98. Among the subtests the lowest reliabilities were in *Ball Activities* and *Walking Around a Circle.* The authors believe that the lower reliabilities of the *Ball Activity* scores were due to the fact that scoring in this instance was more minute and detailed than

TABLE 67

Motor Achievements Scale

Months	*Score*	*Item*
71	4	bouncing large ball, one hand, distance 1
68	6	catching large ball, elbows at side of body, success on 2 or three trials
65	3	bouncing large ball, both hands, distance 3
65	7	throwing small ball, both hands or one hand, distance 7
63	5	throwing large ball, both hands or one hand, distance 5
62	4	descending large ladder, alternate feet, with facility
60	4	hopping on one foot, 10 or more steps
60	3	skipping, alternate feet
57	6	throwing small ball, both hands or one hand, distance 6
56	3	descending large ladder, alternate feet, with caution
55	5	catching small ball, elbows at side of body, no success on one trial
55	4	descending long steps, alternate feet, unsupported
55	3	hopping one foot, 7 to 9 steps
53	4	throwing large ball, both hands or one hand, distance 4
53	4	descending small ladder, alternate feet, with facility
52	5	throwing small ball, both hands or one hand, distance 5
51	3	descending small ladder, alternate feet, with caution
51	5	catching large ball, elbows at side of body, no success or success on one trial
50	4	catching small ball, elbows in front of body, success on 2 or 3 trials
49	4	descending short steps, alternate feet, unsupported
48	3	descending long steps, alternate feet, with support
48	3	descending short steps, alternate feet, with support
47	4	ascending large ladder, alternate feet, with facility
46	2	bouncing large ball, both hands, distance 2
46	3	jumping 28 inches, alone, feet together
46	2	hopping one foot, 4 to 6 steps
45	3	ascending large ladder, alternate feet, with caution
45	3	walking circle, no steps off
44	4	throwing small ball, one hand or both, distance 4
44	4	catching large ball, elbows in front of body, success on 2 or 3 trials
43	1	hopping one foot, 1 to 3 steps
43	2	jumping 28 inches, alone, one foot ahead
43	3	throwing large ball, both hands or one hand, distance 3
43	2	skipping on one foot
42	4	hopping both feet, 10 or more steps
41	4	ascending long steps, alternate feet, unsupported
41	3	hopping two feet, 7 to 9 steps
40	2	hopping two feet, 4 to 6 steps
40	5	bouncing small ball, one hand, distance 2
38	2	descending large ladder, mark time, with facility
38	4	ascending small ladder, alternate feet, with facility
38	3	catching small ball, elbows in front of body, no success or successes on one trial
38	1	skipping and shuffle
38	1	hopping both feet, 1 to 3 steps
37	2	catching small ball, arms straight, success on 2 or 3 trials
37	3	jumping 18 inches, feet together, alone
37	3	walking path, no steps off
36	1	jumping 28 inches, with help
35	3	catching large ball, elbows in front of body, no success or success on one trial
35	2	walking circle, 1 to 3 steps off
34	2	catching large ball, arms straight, success on 2 or 3 trials
34	2	descending long steps, mark time, unsupported
34	3	jumping 12 inches, alone, feet together
34	3	ascending small ladder, alternate feet, with caution
33	2	ascending large ladder, mark time, with facility
33	3	throwing small ball, both hands or one hand, distance 3
33	3	jumping 8 inches, alone, feet together
31	3	ascending long steps, alternate feet, with support
31	4	ascending short steps, alternate feet, unsupported
31	2	jumping 18 inches, alone, one foot ahead
31	2	walking path, 1 to 3 steps off
30	2	throwing large ball, one hand or both hands, distance 2
29	2	throwing small ball, one hand or both hands, distance 2
29	3	ascending short steps, alternate feet, with support
29	2	ascending long steps, mark time, unsupported
28	2	ascending short steps, mark time, unsupported
28	1	walking circle, 4 to 6 steps off
28	1	walking path, 4 to 6 steps off
24	4	bouncing small ball, one hand, distance 1
24	2	ascending short steps, mark time, unsupported
24	1	jumping 18 inches, with help
24	2	jumping 12 inches, alone, one foot ahead
24	1	descending large ladder, mark time, with caution

(McCaskill and Wellman, 1938, p. 148. By permission of *Society for Research in Child Development*.)

in some of the other tests. The boys appeared to be slightly superior to the girls on *Steps and Ladders* and *Ball Activities*, and the girls were a little ahead in *Hopping and Skipping*. The scores discriminated well between 2- and 5-year-old groups but not between 5- and 6-year-old groups because the items are not difficult enough to discriminate among the 6-year-olds who were above average.

An excellent summary of tests of agility and strength is given by Bovard and Cozens (1938). They include norms for anthropometric, cardiac, strength, athletic information, and physical efficiency measures among elementary school, high school, and college groups.

C. Dexterity

Tests which are classified as measures of dexterity, characteristically appraise routine or serial perceptions and movements in terms of speed, accuracy, endurance, and force. In some dexterity tests, such as simple tapping, perceptual discrimination is probably a very minor element, and the score is largely determined by muscle and nerve functions. In more complex tests, such as serial aiming or pursuit tests, the perceptual elements may be more important in success than motor co-ordination. Moreover, there seem to be at least two nearly unrelated types of motor co-ordination:

1. Ballistic movement, in which an extremity is thrown or moved rapidly in some highly automatized pattern, illustrated by simple tapping with a stylus
2. Precision movement, in which voluntary control is continuous, as in serial aiming or tweezer dexterity

Among the simpler ballistic tests is the tapping test described by Whipple (1914) and used in many studies since that time. The apparatus for the tapping test consists of a metal plate two inches square, a metal stylus. and an electric counting device. The number of taps recorded in five 5-second trials, allowing 30 seconds of rest between trials, was used as a score by Patterson *et al.* (1930). Others have used various time limits from 5 to 60 seconds. Maximum retests correlations of about .91 were found by Gates (1928), Muscio (1922), and Greene (1931) under optimum conditions. Greene reported that 10 seconds was the optimum time limit for single trials of college students. Shorter periods seemed to introduce random variations, and longer periods were influenced by individual differences in fatigue. He also found that speed of tapping for 10 seconds with paper and pencil, a group test, correlated .77 with a 10-second trial on the Whipple apparatus.

Another tapping test which seems to measure about the same sort of co-ordination is pressing and releasing a telegraphic key as rapidly as possible. A fairly large number of researches have used tapping keys and automatic counters.

Tapping tests among adults have been reported to show nearly zero correlations with other motor tests and with mental-test scores. Link (1918) found a correlation of .52 between a tapping test and speed in fitting shells into a gauge — an operation which requires quick automatized movements.

Another test described by Whipple (1914) is called the Three-Hole-Aiming Test. This requires the rapid insertion of a stylus into three holes, placed at the corners of an equilateral triangle, as many times as possible in 60 seconds, Figure 58. The final score is the average of five trials.

One of the earlier batteries of tests which emphasized precision of movement was that of Whitman (1925). His series included the following activities:

1. Putting one brass pin in each hole drilled in a large board, each hand, 1 minute
2. Putting three brass pins in each hole, using both hands, 2 minutes
3. Assembling nuts and bolts, 30 seconds
4. Disassembling nuts and bolts, 30 seconds
5. Sorting different colored pegs, 30 seconds
6. Placing pegs of a particular color in order on a peg board, 1 minute

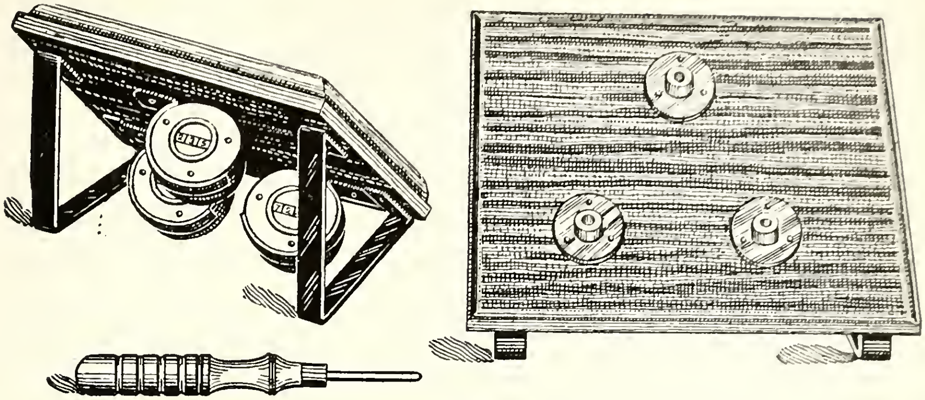


FIGURE 58. **Scott Three-Hole Aiming and Tapping Test**

(Courtesy of the C. H. Stoelting Co., Chicago, Ill.)

Whitman furnished age norms for groups from 7 to 15 years of age. O'Connor (1928) used a test similar to No. 2 and also described a test called Tweezer Dexterity, which requires a person to place metal pins, one at a time, in one hundred holes drilled in a metal plate, using a small pair of tweezers. O'Connor's dexterity tests showed retest reliabilities of approximately .90 among groups of workers. They have been widely used and they have predicted success fairly well in jobs which require delicate assembly of small apparatus.

D. Steadiness

In this group one finds measures of both large and small muscle groups. Paterson *et al.* (1930) described a test of steadiness of large muscle groups called the Body-Balancing Test. The test was to discover how long a person could balance on a three-inch cube of wood on the floor, using just the ball of one foot for support. A simple machine for recording postural steadiness, called an ataxiameter, was described by Miles (1922). It consists of a square balanced board upon which a person stands erect on both feet and with eyes closed. Each corner of the board is connected with an instrument which records its vertical movements.

Great steadiness is indicated by very little total movement. A similar device, called a "wabblemeter," was designed by Moss (1931).

A steadiness test of hand and eye co-ordination was described by Whipple (1914), in which is used a brass plate with nine holes of various diameters, a metal stylus, and an automatic counter. The subject is required to hold the stylus in each hole for 15 seconds while trying to avoid letting it touch the edge. The hand and arm are extended and free from support. The score is the number of contacts which are electrically recorded. The median retest correlation reported by Paterson *et al.* (1930) was .62 for scores on single holes and .76 for total scores amongst 217 boys. Whipple also standardized a thrusting or aiming test with this apparatus.

The existence of a steadiness factor has been suggested by Seashore and Adams (1933), who applied five tests of steadiness to 50 students. The tests were Miles' ataxiameter, Beall and Hall's ataxiagraph, Seashore's modification of Whipple's steadiness test for both "position" and "thrusting," and a rifle steadiness test.[1] The intercorrelations of these tests ranged from .44 to .59, median .48. These correlations are probably indicative of a general factor of steadiness in large muscle co-ordinations.

E. Motor Rhythms Tests

Ability to perform acts at a particular rate is considered to be highly important in many sorts of musical, dancing, and mechanical skills. It is difficult, however, to find a precise definition of rhythm, and a brief inspection of tests of rhythm will convince one that the patterns vary considerably. Subjective analyses usually limit the phenomena to temporal patterns and distinguish between *perception* of rhythmic patterns and their *performance.* Perception of simple rhythmic patterns can be tested by asking persons to tell whether two temporal patterns are the same, or whether one of two intervals is longer than another, as in the Seashore (1919) tests of musical talent, Chapter 14.

---

[1] The six members of the university rifle team scored consistently in the upper quarter of the group and the total scores of the six corresponded almost perfectly with their standing in intercollegiate competition.

The performance of rhythmic patterns doubtless requires good perception, but it also involves muscle and central nervous system elements to a large extent. An illustration of a standardized test of simple motor rhythm is that of Seashore (1928). This requires one to listen to a sequence of four notes that are played over and over again by a phonograph record for a short time. Then at a signal, one is required to keep time with the sequence by pressing a telegraph key. The score, which is automatically recorded electrically, is the number of taps in one minute which fall within .05 seconds of the exact time of the sound.

Van Alstyne and Osborne (1937) adapted this test for small children and also devised a rhythm memory test. They found boys to be slightly inferior to girls. Very small practice effects were noted. Liebold (1936) found that 4-year-olds could not keep time to a metronome at the rate of either once or twice per second, but that one-half of the 5-year-olds and nearly all of the 6-year-old group succeeded very well. A complex test of motor rhythms is illustrated by some intricate dance steps used by Garfiel (1923). She also required persons to attempt nonsynchronized movements of the two hands. Jersild and Bienstock (1935) devised a very accurate measure of stepping and clapping to music. The test scores were secured from moving picture films, by counting the number of frames, taken at intervals of 1/24 of a second, in which the child was in time with the music. Movements of beating time with the hand and walking were counted as in time if they fell within 1/24 of a second of the actual beat. The relative correlations between two tests of 400 beats each were approximately .70, and of 200 beats, .60. Among the older children the figures were somewhat higher. Instructors' ratings were found to be highly untrustworthy when age was held constant. Correlations between hand and foot rhythms were near .82, and between hand rhythms and singing ability, .30.

Rhythms in serial mental work have been studied by Bills (1937). Mental performance was characteristically found to be discontinuous. He reported pauses which became longer and more frequent as the task becomes more difficult and as fatigue increases. Very elaborate studies of rhythmic performance on musical instruments have been reported in the University of Iowa *Studies of the Psychology of Music,* Chap. 14.

F. Batteries of Motor Tests

Several batteries of motor tests have been described which seem to have been assembled, like a general mental test, to sample a fairly wide variety of skills without attempting a careful analysis. Success on such batteries seems to depend, in various and unknown proportions, upon strength, endurance, precision, steadiness, rhythm, perceptual skill, information, and in some cases upon reasoning and planning. Most of the batteries seek to furnish separate norms for their component tests, thereby providing tools for subsequent analyses and refinements.

Two tests which seem to require principally precision, rhythm, ballistic movements, and planning, come from the Minnesota University Laboratory. The Minnesota Rate of Manipulation Test, Ziegler (1934), Figure 3, page 21, shows the speed at which a person can turn over 120 circular blocks which fit loosely in holes in a board. Paterson *et al.* (1930) described a Packing Blocks test which measures the time needed to place 147 one-inch cubes into a wooden box. Scores for one trial of these tests were found to yield low reliabilities, but the sums of three or four trials usually show retest reliabilities of approximately .90.

In Detroit, Baker and Crockett (1929) applied a series of manual dexterity tests to various school and factory workers. This series included screwing nuts on bolts, packing blocks into a box, and laying blocks along a narrow strip of wood.

Another battery which has been widely used is the Stanford Motor Skills Tests, described by Seashore (1928). Six serial dexterity tests were chosen from among more than twenty tests, so that:

1. They were well adapted for use in schools and factories.
2. They took only a small amount of time, less than two hours.
3. The material occupied little space.
4. They were scored automatically.
5. Each test had high retest reliability, .75 to .86.
6. Each test had a low correlation with the Thorndike College Entrance Examination.
7. Each test had a low correlation with amounts of training in typing, practice on a musical instrument, and training in athletics.
8. The intercorrelations were low, mean .25.

The Stanford motor skills battery includes:

1. The Koerth Pursuit Test: requires a person to hold the point of metal stylus on a metal disc, one-half inch in diameter, mounted on a phonograph. The phonograph is set to revolve once per second, making the metal disc follow a circular path about eight inches in diameter. The score is the distance during which the contact is maintained, during 20 seconds (ten trials).
2. Motor Rhythm: described above.
3. Tapping Key: speed of pressing and releasing a telegraphic key for 5-second periods (three trials).
4. Serial Discrimination: speed of reaction to the four numbers, 1, 2, 3, and 4, exposed visually in random order. The reaction is made by pressing the correct one of four keys which correspond to the numbers. Each key is to be pressed and released by a different finger. The score is the number of correct reactions in 2 minutes.
5. Brown Spool-Packer: speed of packing spools in a small box using both hands. Score is the number of spools packed in 3 minutes.
6. Miles' Drill Test: speed of rotating a handle of a small hand drill for 10 seconds (three trials).

V. CORRELATIONAL ANALYSES

A number of correlational analyses which included performance and motor tests have been cited in Chapters 7 and 11. Several typical analyses, planned to evaluate factors in mechanical and motor ability, are described here. Many

TABLE 68

Tests Used in a Study of Motor Ability

	Retest Reliability	*Correlation with Criteria*
1. MENTAL: United States Army Alpha Score	.92	.02
2. MOTOR SPEED:		
a. Tapping speed: Hand metal stylus, 60 seconds	.69	.22
b. Foot speed: Stationary running, 30 seconds	.76	.23
c. Running: 100 yards indoor track	.85	.63
3. MOTOR CO-ORDINATION:		
a. Steadiness: Brass plate and stylus, 10/64-inch hole, contents in 60 seconds	.61	.19
b. Three-hole test: Insert stylus, 100 times	.60	
c. Target throw: Tennis ball, 12 feet, five throws	.07	.20
d. Picking up paper with teeth: Stand holding right toe in left hand, crossing right foot behind the body, then leaning over and grasping a piece of writing paper made to stand on the floor by folding once the long way. Passed if accomplished in 60 seconds with less than three falls		.44
e. Tricks: Difficult hand and foot co-ordination		.29
4. PREFERENCES: A check list of 12 things to do on a June afternoon, 6 of which were athletic games		.31
5. ANATOMICAL:		
a. Height		.02
b. Weight		.23
6. STRENGTH:		
a. Hand dynamometer	.70	.25
b. Back dynamometer	.81	.40
c. Leg dynamometer	.71	.20
d. Chest strength		.22
e. Lung capacity		.28
7. CRITERION: Ratings of 6 judges on motor ability as shown by strong accurate quick movements. Retest reliability of mean ratings, after 16 weeks was	.92	.92

(Garfiel, 1923. Arranged from Tables VI and VII, by permission of the *Archives of Psychology*.)

persons have questioned whether there is a general motor ability which underlies success in all sorts of muscular skills. The almost universal conclusion is that although certain motor skills show high positive intercorrelations, there is no general or elemental motor factor. Perrin (1921) and Muscio (1922) came to this conclusion after applying balancing, tapping, reaction time, strength, and dexterity tests to small groups of adults.

Garfiel (1923) attempted to find an answer to this question by intercorrelating the results of sixteen tests and ratings of 50 college sophomore girls. Using the appraisals listed in Table 68, she found that the correlations with Alpha scores were nearly zero, and that the median of all the 66 intercorrelations of motor tests was .15. Eleven of these were negative, twenty-five correlations lay between zero and .20, and only four were above .40. However, a battery of tests was selected by a multiple-correlation technique which agreed with the criterion with a correlation of .79. The battery included the eight tests shown in Table 69. The first test, running 100 yards, was allotted by this technique a weight equal to approximately twice the weight of all the other tests together. The fourth test, Tricks, was allotted one-third of the weight of the running test. Scores of two of the tests, Steadiness and Leg Dynanometer, were subtracted from the total in order to raise the prediction of the criteria. Garfiel concluded that there was as much evidence for the existence of a general motor factor here, as for the existence of a general mental factor among mental tests. However,

TABLE 69

Prediction of General Motor Ability

	Multiple Correlation
1. Running 100 yards	.63 *
2. Picking up paper with teeth	.70
3. Back dynamometer	.74
4. Tricks: Complex condition	.75
5. Steadiness	.77
7. Tapping	.77
6. Leg dynamometer	.78
8. Hand dynamometer	.79

* Numbers below this one show increases due to adding each test to the previous total.

(Garfiel, 1923, Table IX. By permission of *Archives of Psychology*.)

small and zero correlations among any battery of reliable tests is probably more indicative of the presence of several unrelated factors.

Seashore (1930) selected eight tests of serial motor skills and applied them to 50 college men. He concluded that the marked independence of highly reliable tests argued against the existence of a general motor ability.

A very thorough study of mechanical abilities of high school boys was made by Paterson *et al.* (1930). They measured over 150 boys on seven reliable tests:

1. Minnesota Assembly Boxes A, B, and C
2. Minnesota Spatial Relations Tests A, B, C, and D
3. Minnesota Paper Form Boards A and B
4. Card Sorting
5. Packing Blocks
6. Nine Hole Steadiness
7. Stenquist Mechanical Aptitude Picture Tests

They also evaluated interest, home activities, technical information, and the quality and quantity of articles made in the shop. They concluded that the low intercorrelations among different measures of mechanical ability suggest a high degree of specificity. Mechanical ability as measured by the tests and shop ratings had no relation to test score of intelligence or agility, and very little correspondence with ratings of environment. With tests requiring no mechanical information, they found no mean differences between academic and mechanical or engineering students on either the high school or the college level.

A factorial analysis of motor abilities among 76 high school boys was made by Buxton (1938) using Thurstone's method. He selected tests which:

1. Included both simple and complex behavior
2. Avoided the influence of fatigue
3. Had high interest value
4. Allowed simple directions
5. Included only serial, repetitive action, where no choices were to be made

The tests were:

1. *Steadiness, thrusting:* S thrusts a stylus into holes on a brass plate which became smaller. The score is the number of thrusts which do not touch the plate, made from a standard distance at a constant rate. Ten trials were made at each hole.
2. *Steadiness, stationary:* S holds the stylus in each hole for 10 seconds. The score is the number of contacts with the plate.
3. *Tapping, three discs:* S taps three metal discs in succession as fast as he can. The discs, 2 inches in diameter, are placed at the corners of a 6½-inch equilateral triangle.
4. *Tapping, two bars:* S taps as fast as possible between two vertical metal bars, 2 inches apart, with a stylus mounted on a frame strapped to the forearm to prevent wrist motion.
5. *Tapping, wrist turn:* S turns an aluminum handle as frequently as possible through an arc of 135 degrees.
6. *Packing, cubes:* S fills a low box with 64, 1 3/16-inch cubes as quickly as possible.
7. *Packing, spool:* S fills a tray with spools as frequently as possible in a short time interval.
8. *Rotor, mobility:* S turns the handle of a small hand drill as fast as possible.
9. *Rotor, pursuit:* S tries to keep a stylus on a dime-sized disc which is rotating on a phonograph-like turntable.

The tests were repeated six or eight times, and gains on the four tests which showed the greatest average gain were included with test scores in the statistical analysis. The analysis resulted in six factors, of which two were tentatively named by the author. One factor was called steadiness because it was found significantly only in the two steadiness tests and in them it was very large. The second factor was labeled manipulation. It was found to be very large in the two packing tests. It therefore might be described as hand and eye co-ordination in prehension, short-carry, and placing operations. None of the other four factors showed very large loadings in any tests. One factor was found in tests which involved the larger muscles of the forearm, upper arm, and shoulder. Another factor was found only in the two-bar tapping, where no visual control was needed, and where the arm movements were ballistic, that is, without precise control after being initiated. Another factor was fairly large in wrist-turning tests and another in packing test gains. The author concludes that no general motor factor can be demonstrated here, and that a larger battery of tests would be needed to show the patterns which may exist.

A factorial analysis was reported by Morris (1939) of results of applying the Pintner-Paterson Series, the Sylvester Form Board, the Lincoln Hollow Square, the Witner Cylinder Test, the Dearborn Form Board 3, the Porteus Maze, the Minnesota Paper Form Board, the Henmon-Nelson Intelligence Test, and the Brown Personality Test. From thirty-three scores for each of fifty-six 9-year-old boys, he found three common factors. One factor was identified as *Spatial Thinking,* a second as *Perceptual Discrimination,* and a third, as an *Ability to Discover or to Use a Rule of Procedure.* Few of the factor loadings were large, and the variances of many of the tests were not well accounted for by the loadings from these factors.

A report by Harrell (1939) showed the relationships between thirty-four tests which were applied to ninety-one cotton mill machine-fixers in Georgia. The entire battery, which required about seven hours of work, included three of the Minnesota tests: the Spatial Relations Boards, the Assembly Boxes, and the Paper Form Boards. Among the seven tests from Whitman's Manual Dexterity Series were the nut and bolt assembly, peg boards, and pin boards. Three of Crockett's Manual Dexterity tests were used: screwing nuts on bolts, packing blocks, and laying blocks along a strip. The seven MacQuarrie tests and four of Thurstone's Spatial tests were included to represent paper and pencil techniques. Additional information included age, school grade completed, experience on mechanical jobs, foremen's ratings of competence and mechanical ability, self-ratings of interest, and three of Thurstone's tests of verbal relations.

From a centroid analysis five factors appeared and were tentatively named *Perception-of-Detail, Verbal Relations, Spatial Visualization, Youth-or-Inexperience,* and *Manual Dexterity*. These factors were identified from the tests which had the highest loadings in one factor and very small loadings in other factors. The first three of these factors are similar to those described by Thurstone, and reported in Chapter 11. The *Youth-or-Inexperience* factor resulted from classifications by age and foremen's ratings.

Some of Harrell's findings about commonly used tests are contrary to usual beliefs. Thus, the Minnesota Mechanical Assembly test did not appear to depend upon either manual dexterity or experience on mechanical jobs. On the second trial this test lost most of the *Visualization* factor and became more dependent upon the *Perception-of-Detail* factor. Furthermore, the Minnesota Assembly Test, the Minnesota Spatial Relations, and the Wiggly-Block Assembly, all of which involve the handling of blocks, had practically the same factor patterns as the Thurstone Spatial Relations Tests and Stenquist Mechanical Tests, both of which are limited to paper and pencil situations. These results support the hypothesis that performance tests which do not require great precision or speed of movement may usually be interpreted as tests of perception (comparison of details) or spatial imagery and reasoning. These two factors were again clearly shown to be independent. One demands rapid comparison of objects which are directly perceived, and the other requires that one imagine how things would look if they were put together or rotated.

If we define a pure test as one which is heavily loaded with but one factor, then the purest tests of the *Perception* factor were found to be routine assembly and stripping, and inaccuracy scores on the Minnesota Spatial Relations test. The purest tests of *Visualization* were Thurstone's Spatial Relations, his test of paper folding (Punched Holes), and the MacQuarrie tests of Copying Patterns, Locations, Counting Blocks, and Pursuit. Fairly pure tests of the factor called *Manual Dexterity* were the placing of pins in holes in boards, the assembling of nuts and bolts, and block packing. These tests require rapid repetition of simple acts with some precision of hand and eye co-ordination and some rough ballistic movements of a nearly automatic variety. There is evidence from other studies, particularly Greene (1939), that precision and ballistic movements are quite independent in some situations. Research is needed to indicate whether or not Harrell's *Manual Dexterity* factor may be resolved into separate factors.

13

MEASUREMENT IN FINE ARTS

Design

I. THE NATURE OF ART

Among the fine arts are literature, music, dancing, drama, and the arts of design which include painting, drawing, architecture, and sculpture. Measures of the arts of design will be discussed in this chapter, and those of literature and music in the next.

The nature of art has been the theme of literally hundreds of large volumes in many languages.[1] Two aspects of art are usually distinguished, its form and its content. Form refers both to mode of expression, such as lines, colors, tones, odors, words, and movements, and to patterns, such as rhythm, balance, and composition. The content of an artistic work is the subject or theme which it is thought to display. Psychologically, the content of a picture may vary considerably among observers, depending upon their interpretations. Both content and form of art are usually related to the motives of the artist, and the real or imputed motives are often an important factor in art appreciation.

---

[1] Illustrations of these treatises are: Chandler (1934). Langfeld (1920), Parker (1926). Santayana (1897). and Tolstoy (1899).

Parker (1926, p. 15) sees three sorts of motives which are distinguishable in many forms of art. He writes as follows:

> The deepest stratum is undoubtedly composed of wishes and emotions, of which the artist himself is hardly aware, whose source or object he has forgotten, going back often to his childhood or some other remote epoch of his life, and cropping out in associations and symbolisms, vague and unaccountable. These provide that aura of mystery necessary to the profoundest beauty in art. The significance of music seems to derive almost wholly from this source; and all works of art are akin to music in this way.
>
> But no less important are the particular problems and conflicts, well known to the artist himself, and not difficult for the biographer to recover, which provide the immediate occasion of the work. These constitute a second layer of meaning in a work of art. Sometimes, as in Watteau's pictures, they are freely and obviously expressed, but on the other occasions they are expressed indirectly by dramatization through some story or legend or historical material. These dramatizations make a third and most evident layer of meaning which a work of art may contain.

It follows from this discussion that evaluation of artistic appreciation and works of art will be difficult. Practically, it appears that there are two personal phenomena to be measured: one, aesthetic appreciation, and the other, the ability to produce artistic compositions. These two seem to be somewhat independent, for a person may be unable to produce a work of art, yet he may have a marked aesthetic response to such work. Appreciation and construction in the arts of design will be discussed separately, and then their relationships will be noted.

II. THE NATURE OF APPRECIATION

Aesthetic responses have been variously defined. They are generally described in such phrases as, a lift of the spirit, a moment of inspiration, an overpowering feeling of beauty mixed with joy and pain, or a feeling of perfection. Some have defined aesthetic responses as feelings of satisfaction or pleasantness and unpleasantness, but more writers on art have not wished to include all feelings. They have preferred to define aesthetic responses as feelings aroused by qualities of beauty. There are several types of experiences which have been called beautiful by various authors.

1. Plato and Aristotle connected beauty with *Ideas or Forms,* which they believed to exist independently of man; but which now seem to be carefully thought-out ideals for their own society. Plato excluded poets from his ideal commonwealth.
2. Some writers exclude from art those satisfactions which are called by them intellectual or practical. Thus, Chandler (1934) distinguishes between eating because one is hungry and eating because one likes the taste or style of food. Likewise, the satisfactions of understanding or mastering a problem in mathematics are not considered aesthetic, but the satisfactions of understanding a symphony or painting or landscape are so considered. Discriminations of this sort are controversial and tend to conflict with the other definitions of the aesthetic.
3. Allison (1821), and others before him, emphasized the idea that artistic experiences are satisfactory because they evoked many thrilling or pleasing associations. This definition is so broad as to include any type of experience, and hence it makes aesthetic responses and feelings synonymous.
4. Many writers have suggested that artistic experience is one in which a person feels himself to be a part of the art object. This is called *empathy*. Satisfaction through empathy is gained in part by securing a desired position which takes one out of a mediocre environment of petty responsibilities and irritations. Empathy also includes the satisfactions listed in the next two items.
5. Other experts have stated that smoothness of movement. perception, and thought are essential to artistic responses.

This theory is based on a belief that certain physiological balance, rhythm, and co-ordination are basically satisfactory.

6. A large number of writers have stressed the artistic significance of intellectual unity or coherence. A work of art must give to them an experience of a central theme which is well supported by minor themes. The best artists achieve unity but also include a large variety of materials, emphases, or rhythms.

This array of definitions of aesthetic experiences is confusing. It illustrates how various aspects of experience have been singled out as the basis for a definition and it emphasizes the need of defining terms before using them. A number of empirical studies will be quoted which have tried to show the feelings which persons report when confronted with works of art and the responses which lead to their feelings.

A study illustrating basis of art appreciation was made by Lark Horovitz (1938) using five groups of portraits: those of young women, of young men, of old people, of children, and of families.

The results from 523 children, 44 of whom were in classes for those gifted artistically, show the following trends:

1. The preference for certain pictures waned and for other pictures waxed greater with increasing age (see Figure 59). The younger children attached much importance to what the person in the picture could do, whereas the older ones gave more attention to facial expressions, and indications of moods, sentiments, and individual character traits.

2. The preferences of the girls showed marked differences from those of the boys. Boys, in general, selected pictures of men and older persons, and girls selected women's and children's pictures. At all ages there were strong tendencies to select portraits which depicted ideals, such as Indians, soldiers, — strong adventurous types for boys, and feminine ideals for the girls (see Figure 59).

3. Color was of less significance than the subject matter (see Figure 60) and of more importance than technique, reality, or

emotion with those under the age of approximately 14 years. With those above that age, the order of importance was reversed. The differences between these factors are much larger at 8 years than at 15 years, as shown by per cents of answers.

4. The special class pupils (12 to 15 years) differed from the average youngsters of the same age by showing less interest in subject matter and facial expression, and more in technique and design. Their interests showed no differences of interests between the sexes.

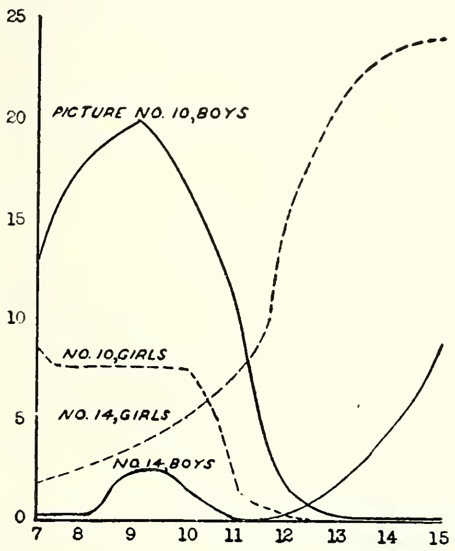


FIGURE 59

Preferences for Two Portraits by Age Groups

Picture No. 10 by K. W. Leighton, *Sun Dance Bustle;* No. 14 by George Buchner, *Erica*

(Lark-Horovitz, 1938, p. 582. By permission of the *Journal of Educational Research.*)

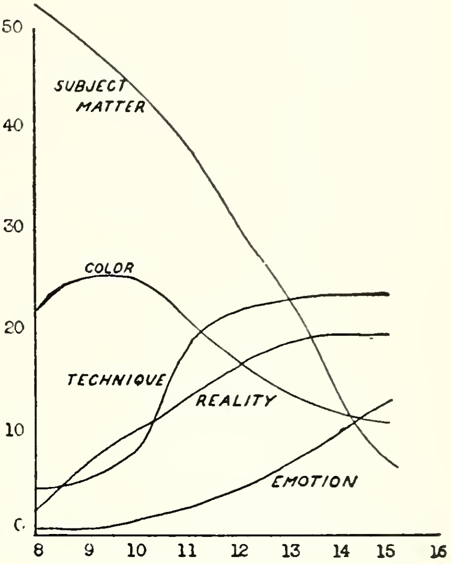


FIGURE 60

Age Variations of Interest in Subject, Color, Reality, Technique, and Expression of Emotion in Portraits

(Lark-Horovitz, 1938, Graph XII. By permission of the *Journal of Educational Psychology.*)

5. The results also showed that approximately 50 per cent of the preferences were given to from 2 to 6 pictures in each age group, and that the same pictures tended to be popular with all age groups. These few pictures had some unusual power of attraction. Either they appealed to many persons for different reasons, or they possessed a single combination of qualities which are universally pleasing.

This study shows that preferences of relatively untrained children were closely associated with interests and with their intellectual comprehension of the total situation.

Further analyses of processes which lead to preferences were reported by four investigators at Iowa working under the direction of Meier. Brighouse (1939) exposed a colored lantern slide of a picture for .26 of a second and recorded the verbal responses of each subject after each exposure. The exposures were continued until the subject had nothing more to add. The responses of 22 children of about 10 years of age, 38 adult nonartists, and 34 artists led to the following conclusions:

1. More than three-quarters of the children and the untrained adults gave judgments of pleasure, indifference, or dislike to the first exposure, and 94 per cent had made judgments by the fourth exposure. These judgments were seldom changed. Among the artistic group only 55 per cent had expressed a feeling reaction by the fifth exposure and these were frequently changed or modified later.
2. Three stages of intellectual evaluations appeared. The first was a quick survey of the objects and the story of the picture, reported in the first three exposures. Children seldom went beyond this stage. The second was an examination of small details. The third, an integration of previously observed features, was seen in all the artists, but in only three of the untrained adults, and in none of the children.
3. The types of responses were further analyzed. Most of the untrained adults and children felt pleased or otherwise toward persons in the picture or their imagined characteristics. The untrained adults reported many more associations than either the children or the artists. The artists reacted almost exclusively to compositional patterns which became clear and complete only after many exposures.

Kellett (1939) presented her subjects with pairs of pictures, one a well-unified painting and the other a less-unified photograph of a very similar situation. Among the artistically unsophisticated she found that pleasantness was not clearly associated with the unity of the stimulus and that the individual observer often introduced new elements or failed to notice those which were presented.

In a further effort to define unity and to analyze responses to paintings, Clair (1939) asked an observer to look at Renoirs' "Le Moulin de la Galette" and then said:

> We're trying to find out what things there are in pictures that interest all sorts of people. I'd like to have you look at this painting (indicating picture) and tell me just what you think of it. Look at it as long as you like, and as you do so, think of it as a composition that has been planned, or organized, or patterned, or designed. Consider all the things the artist had at his disposal — the lines, the darks and lights, the colors, and textures, and shapes, and spaces — tell me all that you can about the ways he has arranged these things to make the composition more pleasing to you.

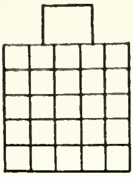
The student's answers were noted on the check list shown in Table 70, which contains spaces for five reports on each of 42 aspects of a picture. Results from 165 high-school students in art courses were compared with those from 200 students with no formal art training. Both groups had similar distributions of items checked, but the art group was more aware of line and motif, and the untrained, of color. Only 65 per cent of the trained group believed that the picture was balanced, compared with 85 per cent of the untrained. Clair believed that the trained group might have come to this conclusion because their training had been largely two-dimensional, whereas the paintings showed a three-dimensional scene. The proportion of persons who "liked" the picture rose from approximately 78 per cent to 83 per cent during the analytical consideration and the oral report to the investigator.

Consistency of preferences was studied by Cahalan (1939) for whom 53 trained and 100 untrained students indicated reasons for preferences on 63 pairs of pictures on two occasions one year apart. The mean of the trained group marked 82 per cent of the choices alike on both occasions, and the untrained, 78 per cent. The persons with the

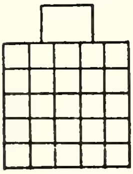
TABLE 70

A Check List of Responses to Paintings

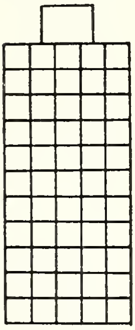
(To be filled in by the examiner)



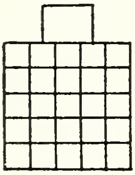
I. LINE ORGANIZATION
1. repetition
2. sequence
3. linear rhythms
4. line tension
5. emotional effect of line



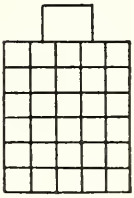
II. VALUE ORGANIZATION
1. variation in size of areas
2. contrast
3. grouping
4. balance of dark and light
5. emotional effect of value arrangement



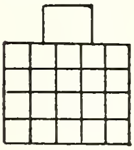
III. COLOR ORGANIZATION
1. repetition
2. sequence
3. grouping
4. variation in size of areas
5. variation in intensity of color
6. variation in hue
7. contrast
8. balance
9. colors used to give movement
10. tonality
11. emotional quality of color



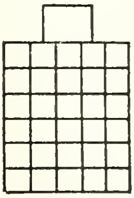
IV. MOTIF ORGANIZATION
1. repetition of shapes
2. sequence of shapes
3. variation in size of shapes
4. variation in shape of areas
5. emotional effect of shapes



V TEXTURE ORGANIZATION
1. repetition
2. sequence
3. contrast (plain and texture, etc.)
4. balance of texture
5. texture used to give movement
6. emotional effect of texture



VI. VOLUME ORGANIZATION
1. balance of weight three dimensionally
2. movement through arrangement of volume
3. tension of volumes
4. emotional effect of volume tensions



VII. SPACE ORGANIZATION
1. planes parallel to the picture-plane
2. planes at an angle to the picture-plane
3. tension of planes
4. space rhythms
5. balance of tensions
6. realization of "picture space"

(Clair, p. 55, 1939. By permission of the Editor, *Psychological Monographs.*)

highest score, according to experts, were more consistent than those with lower scores.

These studies of the bases of artistic appreciation give evidence that the amount of satisfaction reported after viewing a picture varies according to number of apparently independent responses. Detailed perceptions, familiar personal associations, and ideals of artistic unity were the most frequently reported variables. Physiological responses may have been a large factor in some cases. The standard tests discussed below have usually ignored the basis of feeling and asked persons simply to indicate their preferences.

A. Tests of Preferences for Pictures

In the field of appreciation of graphic art, two tests have been developed and widely used: the McAdory Art Test (1929) and the Meier-Seashore Art-Judgment Test (1930). Both of these require that a choice be made between samples which are presented.

The McAdory Art Test consists of 72 plates each of which contains four small variations of one picture, Figure 9, page 35. The original drawings were taken from current art and trade magazines, and the variations involve changes in proportions, intensity, and color, which are described on the record sheet. The subject records his order of preference for each plate. One point is given for each picture ranked according to a key which represents the judgments of 100 experts. The experts included artists, architects, art teachers and critics, art buyers, and lay critics. All the keys used were agreed upon by at least 64 per cent of the experts. Separate scores can be had for the total test and also for six subdivisions, which are:

1. Furniture and utensils
2. Textiles and clothing
3. Architecture and related arts
4. Shape and line arrangements
5. Massing of dark and light
6. Color schemes

The Meier-Seashore Test consists of 125 small uncolored pairs of pictures, Figure 61, and a record sheet. One member of each pair of pictures is a reproduction of a recognized masterpiece chosen from landscapes, pottery, portraits, oriental drawings, woodcuts, murals, and medallions. The other member of each pair is a reproduction of the masterpiece altered in some respect. As these alterations are noted on the record sheet, the examinee's attention is called to them. They include changes of position of an outstanding object, of background, of distribution of light and shade, of horizon line, of perspective, of quality of line, and of the use of curves. Although there are no time limits for the administration of this test, it usually takes from forty to sixty minutes. The score is the number of choices which correspond to a key. The key was made from the consensus of opinion of various artists, sculptors, directors of art training, and art teachers. The items are arranged roughly in order of difficulty of discrimination by experts.

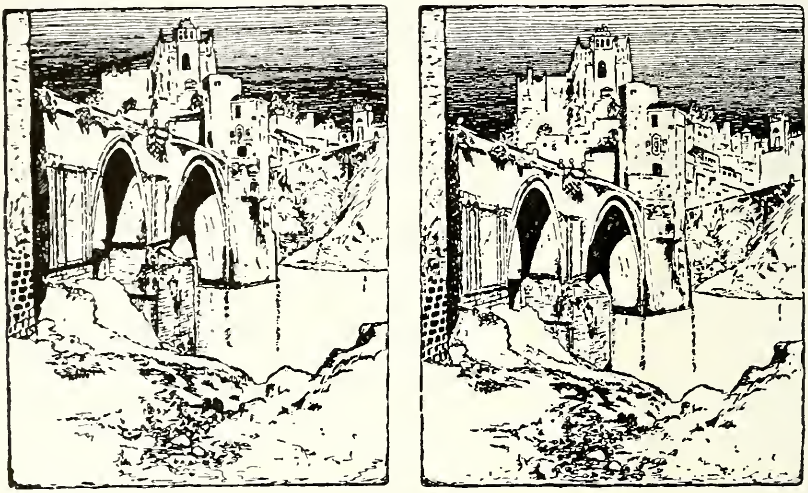


FIGURE 61. **Art Judgment Test**

San Juan Bridge. The relative proportions of the bridge and the city have been altered.

(Meier-Seashore, 1930, No. 78. By permission of the Bureau of Educational Research and Service, University of Iowa.)

From the application of the McAdory and Meier-Seashore Tests by various investigators to fairly large numbers of persons, it has been found that:

1. Retest correlations ranged from .71 to .93, and odd-even correlations of .59 and .65 were found (McAdory, 1929, and Meier, 1930).
2. The Meier-Seashore and McAdory Tests correlated with each other .37 in small groups of college or art students (Carroll and Eurich, 1932, and Wallis, 1930).
3. Age and grade norms show improvement in the Meier-Seashore Tests from nearly chance mean success in the eighth grade to the 87 out of 125 for art teachers. Norms for the McAdory Test have been secured from the third grade up.
4. Women exceed men in mean scores on both tests by small but fairly significant amounts (Eurich and Carroll, 1931).
5. Artists and students with artistic training have average scores which are significantly higher than similar college students without such training.
6. Studies of Mexican children (Stolz and Manuel, 1931) showed them to have about the same scores as non-Mexican children of the same general environments and age, both groups having very low mean scores. Steggerda (1936) found 300 Navaja Indian children, ages 11 to 16, to be far below the McAdory norms of New York City whites, and concluded that the test does not reveal the artistic ability of these Indians.
7. Correlations between scores on intelligence tests and art appreciation tests are usually between .07 and .26 (Tiebout and Meier, 1936).
8. Correlations between art and appreciation tests and Bernreuter's introversion, submissiveness, and emotional stability were all nearly zero, among 218 students (Carroll, 1932).

Attention should also be called to a test by Whitford (1920), a University of Wisconsin Test described by Drought (1929), and a test by Christensen and Karwoski (1925). None of these tests have been widely used but they illustrate interesting approaches.

The Whitford Test included fourteen preferences to be made in 15 minutes. Appropriateness of form, line, proportion, rhythm, color, and perspective were to be judged.

Tests of 1,000 Chicago school children showed that the average fourth grade in "superior" schools exceeded the scores of average eighth grade pupils in "medium" schools.

The University of Wisconsin Test features three aspects of art: unity, proportion, and fitness. Nine plates were prepared each with five small pictures, a "perfect" picture and four variations. Tests of high-school and elementary grades showed progress from preferences of incongruous pictures in the early grades toward preferences for "good art" at the later ages.

Christensen's Test included 105 plates with four colorless pictures to a plate. Separate scores were given for five parts:

1. Paintings of groups of persons
2. Pictures of an individual
3. Architecture and sculpture
4. Industrial arts
5. Designs intended to illustrate abstract art

Bird (1932) believed that these tests indicate conformity to conventional or commonly experienced products rather than true appreciation of an aesthetic sort. Pintner (1918) and Berliner (1918) found that conformity to conventional pictorial representation was well developed at the age of 7 and was almost completely developed at 10, since the ranking of pictures by 10-year-olds was similar to that of average adults.

Voss (1936) found that children in the second, third, fourth, and fifth grades could be taught abstract principles of art composition by having differences between a poor and a good composition pointed out to them. As aesthetic judgment improved, pictures were comprehended less often as substitutes for objects and more often as representations of ideals or moods.

Korpeth-Tippel (1935), working in Vienna, reported that below the age of 3 years the predominant attitude toward pictures is nonaesthetic. Between the fourth and tenth years aesthetic preferences developed gradually, and the eleventh year shows a marked transition. Nearly all the 14-year-olds select the more artistic pictures of the pairs presented.

There is no way of indicating from most of these tests what the nature of aesthetic experience is. All one can say is that under the test conditions certain preferences were indicated. Whether or not a particular person responds to one or several aspects of a picture cannot be known unless we go further and try to find the reasons for his preferences.

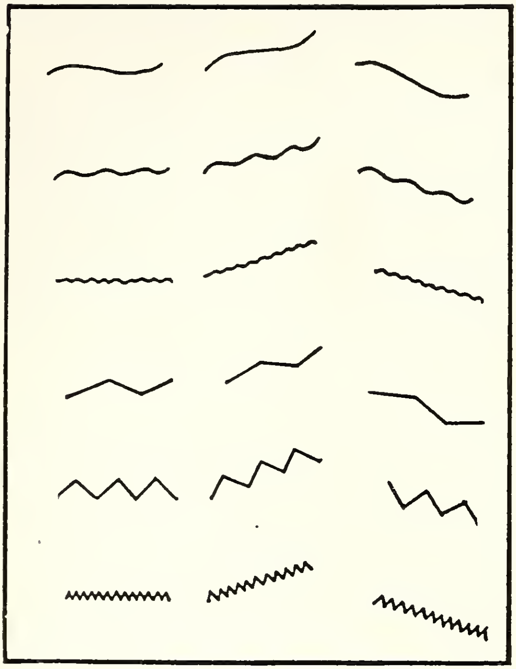
B. Responses to Lines

Closely allied to the composition of a whole design is the aesthetic evaluation of small fragments of a design. Three studies illustrate this approach, using both matching and reproduction techniques. Poffenberger and Barrows (1924) asked 500 adults to look at 18 printed lines, and then to choose, from a list of words, one which described the way each line made them feel (see Figure 62). A marked tendency was found for a slow descending curve to be indicative of *sad, lazy,* and *weak;* slow horizontal curves, *gentle* and *quiet;* medium rising curves, *merry* and *playful;* rapid rising angles, *agitating, furious,* and *powerful.* A conclusion was reached that "direction of line was generally most important, rhythm next, and form least in this particular study."

In the Guilford's (1931) study, 24 adjectives were selected and then printed with a blank space. The instructions were:

> In the space below you will find a list of adjectives. Take each in turn. You are to think of its meaning and then draw a single line which best expresses the meaning that it conveys to you. Your lines will be graded in their general form, directness, and heaviness. You will have ten minutes in which to complete this test.

The adjectives were *sad, forceful, dead, earnest, playful, tranquil, lively, cruel, joyous, quiet, grave, gentle, lazy, fiery, furious, jolly, hard, agitating, angry, faint, delicate, idle, sorrowful,* and *strong.* The results were scored to show the con-



LIST OF WORDS

1. Sad, melancholy, mournful, doleful, sorrowful
2. Quiet, calm, tranquil, serene
3. Lazy, indolent, idle
4. Merry, cheerful, gay, jolly, joyous
5. Agitated, excited, fiery, brisk, vivacious, lively
6. Furious, angry, cross, vexed, enraged
7. Dead, dull
8. Playful
9. Weak, feeble, faint, delicate
10. Gentle, mild
11. Harsh, hard, cruel
12. Serious, solemn, grave, earnest
13. Powerful, forceful, strong

Note: One word is to be chosen from the list to indicate how each line makes you feel.

FIGURE 62. **Feeling Value of Lines**

(Poffenberger and Barrows, 1924. By permission of the *Journal of Applied Psychology*.)

formity of an individual to the tendencies of a group of 144 students, 55 of whom were studying design. For instance, the word *idle* was represented as follows:

Form		*Direction*		*Line*	
Angular	9	Horizontal	110	Heavy	4
Curve	57	Upward	9	Medium	56
Wave	28	Down	25	Light	84
Straight	50				

This tabulation shows that most students drew a light horizontal, curved, or straight line for the meaning of the word *idle*. Students who made such lines were given more credit than others. Wider deviations were given less credit than smaller deviations.

The scores showed split-half reliabilities of approximately .65 among two groups, each containing 50 students in design. The tests scores showed correlations of .84 and .83 with teachers' estimates of originality and fertility in design when both the tests and estimates were corrected for attenuations, but only .58 and .65, when uncorrected. The correction for attenuation theoretically removes chance errors of measurement (see Chapter 4). The actual predictive value of the test, however, is shown by uncorrected correlations. The uncorrected correlations between test scores and art grades were .48 and .58. These are fairly high for this sort of test. The art test scores correlated .18 with the United States Army Alpha for 69 liberal art students, and .08 with the American Council on Education College Psychological Examination for Freshmen among 48 women students. Apparently the line-drawing test involves abilities not called for in verbal mental tests.

Walton (1936) used both of the Guilford and the Poffenberger and Barrows techniques. Tests of matching lines with words correlated with giving words to represent lines in the neighborhood of .51. Self-correlations between two trials of either test ranged between .26 and .70 for small groups of children. The correlations between the test of drawing lines to represent meanings of words and other art tests were all approximately .31. Grades in school, grades in art, intelligence test scores, test of discrimination of color harmony, and sex appeared to be unrelated to tests of matching lines and word meanings.

C. Responses to Color

Aesthetic responses to single colors have been studied with the hope of finding basic or marked tendencies of feeling or association. The evidence seems to show that when texture and brightness are eliminated, the response to color is largely a learned one.

Early experiments are well illustrated by Stefanescu-Goanga

(1911), who asked 6 subjects to give introspective reports of the effect of each of a series of colors. The number of times each color was found to be exciting, soothing, pleasing, and displeasing was reported. In general, red, orange, and yellow were stirring, warm and cheerful; green was quieting; blue, indigo, and violet were depressing, serious, cold, and sad; purple was exciting and dignified. The feelings were thought to be aroused before the associations of specific sorts.

Bullough (1910) classified responses to colors into four main groups:

1. Objective: such as judgments of hue, saturation, and luminosity
2. Physiological responses: one's awareness of bodily feelings, such as excitement, weight, and warmth
3. Associations of specific sorts: moon, water, fire, and medicine
4. Character: i.e., imputing to a color a particular character or set of personal traits

He reported that in most of his observers one class of response predominated and that each type of response was usually incompatible with the others. The objective responses seemed typically intellectual; the physiological, merely adjustive; and the associative; variable or intellectual. The "character" responses were considered to be the most aesthetic because they "objectified emotion."

The findings of Bullough were not supported by the monumental work of Van Allesh (1925), which reported twenty years of observations. He found that tints and shades of all colors, when background and suggestion were avoided, evoked no typical responses, and that all observers were inconsistent in some manner. He found so little consistency in introspective responses to short exposures (7 seconds) that he avoided general conclusions. The most frequent effects reported were activity and passivity, cheerfulness, sadness, earnestness, loudness, softness, friendliness, hostility, warmth, and coolness. He also described the appearance of color as a film, a surface, or a space. He sought to have his observers see only film colors, because the other aspects seemed to arouse additional associations.

He also found disagreements and self-contradictions in preferences for pairs of colors. The responses to color pairs added ideas of harmony, based somewhat on associations with single colors. Thus, a dull brown and a bright orange seemed to represent to one observer severity and pride which did not belong together. Verbal suggestion of solemnity, grief, friendliness, joy, poverty, gave color associations which were not otherwise made. Every color and color pair could be made, by suggestions, to seem commonplace, and nearly all colors could be associated with both the tragic and comic. Color pairs were as susceptible to suggestions as single colors.

Children's responses to color have been studied by about twenty investigators using various materials and techniques. The materials were usually colored papers which were not, of course, pure and which have a definite texture as well as a color. Generalizations from such materials should not be made freely, but the following conclusions are fairly well agreed upon:

1. Infants of 70 days of age discriminate between red, orange, green, and blue green when brightness differences are eliminated (Chase, 1937).
2. Four colors are clearly differentiated by the age of 15 months (Staples, 1932).
3. Children between 8 and 24 months are most effectively stimulated by red, next yellow, blue and green in order (Holden and Bosse, 1900).
4. Groups of 4- and 5-year-olds stated that orange was the favorite color; pink, second; and red, third.

Williams (1933) and Walton (1933) found that high scores in sensitivity to color harmony, as shown by the selection of a colored scarf to go with a colored dress on a doll, were found as early as the fourth year of age, but that age group averages did not exceed chance until the eighth year. The color test scores indicated the amounts of agreement with judgments of experts. There was a gradual increase in scores to the twelth year. College students did much better than 12-year-olds. Intelligence test scores, teachers' ratings of artistic ability, and scores from art tests based on form discrimination were not closely related to scores on this test of color harmony.

D. Discussion of Tests

From the description of tests of artistic appreciation just given, it appears that:

1. Total scores on group tests of 100 items or more give fairly stable group means.
2. Total scores are usually not reliable enough for appraisal of individuals, except when mature students of art have been tested on discrimination of pattern.
3. Low correlations between various subtests in art appreciation batteries indicate the probable existence of several independent patterns of response. This situation results in ambiguity of interpretation of scores, for a given score can seldom be said to refer to a particular pattern of response.

These conclusions point to the need of more analytical appraisals which would show a person his profile of independent skills. A very fascinating field of research lies in this direction. From a survey of existing work, it seems probable to the writer that the following will eventually appear to be among the independent aspects of picture appreciation:

1. Discrimination of mood of isolated lines
2. Discrimination of coherence of (1) lines, (2) areas, and (3) three-dimensional spaces
3. Discrimination of mood of colors
4. Discrimination of minute variations of theme

5. Other specific associations of personal memories, desires, prejudices or fears, as shown in part by the Rorschach technique, Chapter 17
6. Feelings of satisfaction associated with the activities listed immediately above

With this tentative analysis of aspects of appreciation, let us turn to the consideration of activities involved in artistic composition.

III. THE NATURE OF ARTISTIC COMPOSITION

The psychological processes actually used in the creation of a design have generally been reported only in fragmentary anecdotal fashion. Recently, however, three studies which attempted a systematic observation of artists or others at work have appeared. Grippen (1933) found that creative artistic imagination was rarely demonstrated below the age of 5, as shown by spontaneous drawings. By watching artistically talented 6-year-olds, he discerned the following seven types of imaginative development:

1. Revision of a single memory image
2. Organization on the nature of a composite from several images, usually related
3. Improvisation of a theme, resembling the source or sources, from a number of images
4. Selection of various elements of aesthetic interest, to which other elements may be added, all based upon a single memorial or sensory experience
5. Compositional expressions arising as a reaction from a single memory touching upon some more or less strong emotional experience
6. Effective expressions appearing in appropriate compositional setting from a single vivid aspect of a larger experience residing in the child as a memorial experience
7. Fusion of compositional elements or aspects into a composition of high character, from a continuing experience over a limited time interval

A comparison of 67 drawings of talented with 79 drawings of nontalented children by Grippen showed large differences as follows (arranged from p. 80):

Proportion of Drawings	*Talented*	*Nontalented*
Exhibiting composition skill	90 per cent	5 per cent
Number of verbal self-criticisms	77	4
Mean number of relevant verbal comments	64.3	72
Mean time per drawing	15 min.	3 min.
Mean number of colors used	3 plus	1 plus

These figures indicate much greater interest and activity, both mental and physical, and greater perseverance on the part of the talented children. The relative importance of motivation and ability cannot be evaluated from this study.

Patrick (1937) recorded the verbal expressions of 50 artists and 50 unpracticed sketchers who agreed to talk about their work as they worked. He concluded that both groups revealed the same four stages, which were also found in poetic creation — namely, preparation, incubation, illumination, and verification.

Lark-Horovitz (1936) studied the drawing of untrained adults, 74 men and 96 women, whose average age was about 25. The following objects were drawn from memory: violin, chair, church, duck, horse, man, woman, child, flower, automobile, and a country road. She followed Kerschensteiner's (1905) analysis of three general stages of development:

1. Schematic representation typical of children
2. Presentation true to appearance in two dimensions
3. Perspective drawing

These untrained adults made drawings much like children from 6 to 10 years of age; most of them were of the schematic type. A questionnaire showed that most of them had tried to reconstruct the picture from memory on some logical basis but were unable to disentangle various schemata. Manual ability was thought to be of small importance, but failure to grow in graphic presentation seemed due to lack of training in visual discrimination and synthesis.

To some extent the processes in artistic composition are also indicated by studies of development of drawing ability.

A. Development of Drawing Ability

An examination of the extensive literature on children's drawing[2] shows a widespread belief in the existence of several stages of development. There is fairly general agreement that the preliminary stage is one in which scribbling marks are made with no attempt at representation. Occasional similarities between objects and drawings may be noticed by the child. In the next stage, important details of a person or house may be drawn as they are remembered, with only slight success in putting them together. Thus, Figure 29 shows a man whose hands and legs are attached to his head in the absence of any body. This stage is usually one where motion or use is depicted. In a later stage the idea of the whole tends to dominate, so that the details have their proper relative size and position. Correct perspective is a still later technique. In older children more attention is given to style and fineness of detail and composition. A few specific studies state the ages of development for certain types of discrimination and compositions.

Daniels (1933) found that preschool children, ages 2 to 5 years, showed marked preferences for balanced block designs, but that preference for balanced designs was not correlated either with ability to reproduce the design with blocks or with Stanford-Binet scores. Whorley (1933) found that unified compositions of toy trees and of furniture on standard backgrounds were rarely made before the age of 4. Such unity increased gradually to the age of 10. The test scores involving outdoor arrangement correlated with scores of indoor arrangements between .29 and .45 for various small groups. The investigator felt that "fitness" may have been a more important factor in the indoor model than in the outdoor model.

---

[2] Ayer (1916), Goodenough (1928), Tomlinson (1934), Meier (1935), and Anastasi and Foley (1936).

Saunders (1936) found that two years of intensive training showed radical changes in art ability among children in the first four grades. The improvement bore a direct relation to the amount of instruction and to the initial degree of artistic inferiority. Unfavorable home conditions, use of improper materials, lack of motivation, and lack of sensitivity to elements of artistic quality were found to be important deterrents to the development of drawing ability.

These reports all agree that artistic composition in the field of design is very much improved by familiarity with types of composition and their fine points. Interest, amount of mental activity, perseverance, and length of special training were very effective in improving artistic compositions.

B. Tests of Composition and Representation

The ability to compose an artistic work is not easy to distinguish from the ability to draw mere representations of objects.

Since there are no sharp lines between an artistic picture and one not so artistic, scales of values have been established on the basis of agreement of judges by Thorndike (1916), McCarty (1924), Kline and Carey (1922 and 1923), and Tiebout (1936). The Thorndike scale is a series of 34 drawings, each of different subjects, arranged in order of excellence by 60 artists and 60 art teachers. McCarty's scale consists of three series of 34 drawings each, one of persons, one of horses, and one of landscapes. Each series is divided into 9 steps in a scale based on ratings of 60 judges. The drawings were collected from children, ages 4 to 8 years. The two Kline-Carey scales are intended to measure two different aspects of composition: representation and design. The first consists of 14 graded samples of each of the following: house, tree, rabbit, and a figure in action drawn to illustrate a short story. The second includes 10 or 11 graded samples of each of the following: illustration, poster, border, and structural design. Each sample is given a numerical value based on order-of-merit selections of 152 judges who were persons with considerable art training. Tiebout (1936) introduced color, and followed the plan of having children illustrate short stories or parts thereof. Paintings were made independently by each child using tempera paints under standard conditions. One hundred children in each grade from the first to the seventh inclusive contributed three drawings each. Four judges made preliminary selections of 30 paintings for each grade and ranked these in five piles according to artistic quality. Fourteen experts then arranged these selected paintings. They were advised to give consideration to "attainment of rhythm, balance, unity, and other aesthetic qualities, rather than to technique or realistic representation." A few well-qualified judges were considered to be more effective than a larger number of less experienced judges in the establishment of artistic value. In its final form the scale has from 11 to 16 paintings for each grade. The judges' agreement upon the relative value of the scaled pictures was high; the correlations were .94 and .93 between average ratings of half of the judges with the other half. Two trained workers using these scales for classifying original drawings showed correlations of approximately .79 and .73 with the average of two other-trained workers judging 100 paintings in each grade. No figures were given for consistency of individual judges.

A novel method of appraising compositional ability, which was designed to be free from technical skill, was described by McCloy (1939). The materials used were a control board which varied the intensity and color of a screen on a small stage, a number of small clay statues and objects, and 25 landscapes which could be used as backgrounds. After becoming familiar with the materials, subjects were asked to arrange sets of objects "until you get the effect you like best."

Colored photographs of the preferred arrangements were appraised by three selected judges according to a uniform scheme in which each of the following items were rated on a ten-point scale:

1. Arrangement and use of light and shade
2. Arrangement of figures
3. Background selection
4. Originality
    - *a.* Compositional arrangement
    - *b.* Emotional interpretation
5. Color
    - *a.* Harmony
    - *b.* Appropriateness in background

After testing 20 subjects of various ages and with various amounts of training, McCloy reached the conclusion that creative ability under these conditions bore no relationship to age after 12 years, amount of time used, and previous artistic training. At one extreme the various subjects exhibited arrangements which seemed entirely accidental, and at the other, arrangements which followed definite ideals formed after first seeing the clay forms.

A scale of drawing was constructed by Goodenough (1928) for the purpose of indicating intelligence rather than artistic merit (see Chapter 8). An attempt was made entirely to disregard artistic merit. Nearly 4,000 drawings of a man by children of various age groups were inspected to find what changes took place between successive ages in accuracy of representation. She distinguished eight steps of psychological growth, which are very much like those listed by Grippen (1933) for imaginative development:

1. Seeing a resemblance between pictures and objects
2. Noting the parts to be drawn
3. Selecting the most essential parts
4. Noting relative position
5. Noting relative proportions
6. Representing parts with simplified outlines
7. Co-ordination of hand and eye in drawing
8. The addition of new features as the concept develops

Goodenough's scale consists of verbal descriptions of 52 attributes of a drawing of man, each of which is given one point. Forty illustrative drawings are also printed (see Figure 6). Total points may be changed into a mental age. The points include the presence of parts, their attachment, proportion, and details. Motor co-ordination of the child is scored by the absence of unintentional irregularities. Perspective and figures drawn in profile receive extra credit. This method of scoring drawings undoubtedly gives credit to some unartistic details, but qualities of correct proportion, unity, and fitness, stressed by nearly all the analyses of art, are also given credit. This work shows a fairly clear overlapping of factors in scales of artistic and representative ability in drawing. It would be interesting to have the same drawings ranked for both values, in order to see their closest relationship. No study of this sort has come to hand.

IV. INTELLIGENCE AND ART TESTS

The relationship between intelligence and artistic ability is not clearly defined or measured by published studies.[3] Definitions of both types of ability are characterized by vagueness and lack of wide acceptance. Tests designed to measure these abilities are fairly reliable, but as yet they have failed to give careful analyses of the patterns themselves. Tests of artistic composition usually fail to distinguish between creative and copying processes, and tests of intellect fail to distinguish among such abilities as rote memory, perception, and reasoning. Furthermore, in research to determine the relationship of artistic and intellectual abilities the groups of persons studied have usually been small and rather narrowly selected. The correlations quoted, therefore, are not readily comparable, and differences which are reported do not necessarily indicate differences in essential facts but perhaps merely insufficient samplings. A few typical results are given.

---

[3] Summarized by Bird (1930) and Tiebout and Meier (1936).

Goodenough (1931) reported a correlation of .74 between her drawing scale of intelligence and the Stanford-Binet IQ scores, using 334 children between the ages of 3 and 11 years. Slightly lower correlations were found for these same children when smaller single-age groups were used. Bird (1932) and Tiebout and Meier (1936) found correlations of approximately .43 between IQ's from the Goodenough Draw-a-Man Test and IQ's from two group tests, the Kuhlmann-Anderson and the Dearborn, using 100 or more pupils per grade.

These same investigators found that correlations between the Draw-a-Man Test IQ's and Tiebout's score for artistic drawing were in the neighborhood of .35 for three groups of 100 children each, in the first three grades. In the fourth grade the correlation dropped to .18. Bird's (1930) score for drawing correlated .49 with Goodenough IQ's in a group of 248, six to nine years old. Manuel and Hughes (1932) report correlations still higher, .63 to .86, between Goodenough IQ's and their ratings of quality of drawings in 836 Mexican children in grades I to IV.

Tiebout and Meier (1936) found that correlation between measures of artistic composition and verbal tests of intellect are usually higher in Grade I than in later grades. The Kuhlman-Anderson IQ's and Tiebout scores showed correlations of .35 or .40 in Grades I, II, and III, and nearly zero correlations in the higher grades. No significant correlations of this sort have appeared for adult groups. High school pupils considered to be artistically superior were only slightly above the average Kuhlman-Anderson IQ's. Fifty artists selected as the most outstanding among 5,500 names listed in the *Biography of American Artists* made an average IQ of 118 on the Otis Self-Administering Test, and showed their largest numbers of errors in handling number concepts. Intelligence test scores and rating as an artist showed a zero correlation in this group.

These figures show that persons who produce artistic works are slightly superior in verbal intelligence test scores but that artistic ability is not dependent upon such intellectual capacity.

V. ANALYTICAL STUDIES OF ARTISTIC ABILITY

Four studies should be mentioned because they attempt to appraise skills which were thought to be component elements of artistic composition. Knauber and Pressey (1927) included separate measures of 16 skills, which were judged from drawings or completions of drawings. Their study embraced the following abilities:

1. Memory for designs; long and short time
2. Observation
3. Accuracy
4. Imagination
5. Creative imagination
6. Analyzing ability
7. Ability to visualize
8. Design sensitivity

Lewerenz (1927) prepared nine tests designed to measure basic art abilities. In these tests, which made frequent use of multiple-choice and completion items, he included the following procedures:

1. Preferences for design: Choose between four variations of one theme (14 items of increasing complexity).
2. Originality of line drawing: Draw lines between printed dots to make a picture (10 items).
3. Indicate omission of shadows in ten drawings.
4. Vocabulary of materials, processes, drawing terms, and pictures (50 pairs of words).
5. Immediate memory span: Reproduce part of a picture of a vase from memory.

6, 7, and 8. Indicate errors in pictures of cylindrical, parallel, and angular perspectives.

9. Color matching test: Six key colors are to be matched with 46 variations in hue and shade.

Both of these batteries stress accuracy of observation and memory for details of proportion, shading, perspective, as well as preferences for designs and pictures. This sort of battery affords a means of comparing the relationship between the various subtests and eventually showing by a factorial analysis the basic patterns in artistic drawing. Both of these batteries have subtests which are probably too short for the analysis of factors or the construction of individual profiles. However, the method is sound and a fascinating field of research is open. A correlation between total scores of these two batteries of .64 was found in a group of 64 art students. Correlation between the total Lewerenz test scores and scores on the McAdory and the Meier-Seashore tests were approximately .53. Simple correlations of this sort give no clear basis of analysis, but they suggest that the tests are measuring either the same processes or related processes to a marked extent.

Tiebout (1933) and Dreps (1933) compared scores on various tests of motor co-ordination, observation, discrimination, and memory with ratings of artistic ability. They found that the average scores of small groups of children and adults who were rated as artistically superior exceeded significantly the average scores of similar groups rated as inferior in the following tests:

1. Completeness and accuracy of visual observation (Heilbronner and Lewerenz tests)
2. Recall of observed material after ten days or six-month intervals (Fernald)
3. Uniqueness of interpretation of ink blots (Knox)
4. Originality of line drawing (Lewerenz)
5. Form discrimination
6. Feature discrimination (Greene)
7. IQ's
8. Aesthetic judgment

Small differences of doubtful significance were found on tests of:

1. Recognition memory, immediate
2. Completion of a drawing from memory
3. Visual imagery (Griffitts)
4. Neurotic tendencies (Pressey X–O)

No significant differences were found in tests of:

1. Hand and eye co-ordination (Greene, Whipple)
2. Steadiness of movement (Wellman, Whipple)
3. Color matching (Lewerenz)

Carroll (1932) found very small and unreliable correlations between ability to appreciate or create art and introversion and emotional instability. Artists as a group showed no more instability than a group of students of statistics.

Meier (1939) has recently completed the direction of an elaborate ten-year study of artistic ability which included research by twenty coworkers as well as himself. He described six patterns which he believed to be important in graphic arts:

1. Manual skill: This ability is regarded as fine hand and eye co-ordination which can be noted at early ages.
2. Energy output: This is shown by unusual concentration on a task for long periods.
3. Intelligence: The usual IQ test scores are above average, with more success, however, in parts of the test which have to do with visualizing and speed of perceiving than with number and technical vocabularies.
4. Perceptual facility: By this is meant the ability to observe and recall sensory experiences.
5. Creative imagination: This is defined as an ability to organize vivid sense impressions into a "work having some degree of aesthetic character."
6. Aesthetic judgment: This is considered to be the most important factor of artistic competence. It is defined as ability to recognize unity of composition, and it is measured by the Meier-Seashore Art-Judgment Test.

These six items were not considered to be mutually exclusive but were general terms descriptive of complex and interrelated patterns. Meier believed that the first three factors are primarily inherited through a line of ancestors, and that the last three are limited very definitely by inheritance. He pointed out that future analyses of artistic ability will probably indicate that elemental functions similar to those described by Thurstone (1938) underlie these six patterns. To the writer, the first four patterns seem to be important in any sort of craft or occupation where spatial factors are important. The last two factors seem to be found principally in artists, and hence to distinguish them from other people.

Any discussion of artistic ability would be incomplete without some mention of the analyses of motives which drive artists to their work. Such analysis has been attempted by psychoanalysts, and their work is fruitful and challenging. As yet, however, it has resulted in no standardized tests or ratings which can be included in a chapter on measurement.

14

MEASUREMENT IN FINE ARTS
Literature and Music

The classification of literary or musical activities according to merit is hazardous if one tries to be extremely precise. These terms are customarily used with such a variety of connotations that the task of careful analysis and subsequent measurement is very challenging. Few examiners have as yet accepted the challenge, but the last twenty years have witnessed some progress in the development of standardized testing instruments which are worth noting.

I. MEASUREMENT IN LITERATURE

Although literature is assuredly one of the oldest of the arts, attempts to construct standard measuring instruments in this field are among the most recent. It is convenient to classify these instruments into three groups: tests of discrimination, of appreciation, and of composition.

A. Tests of Discrimination

There are a number of language elements which are certainly basic to literary experiences but which probably have little, if any, relationship to artistic appreciation. Such elements of language include recognition of spoken or written words and knowledge of the meaning of grammatical forms and punctuation. Tests of discrimination of these language elements have been described in Chapter 9.

Another group of tests, designed to measure discrimination of literary style, are illustrated by two tests from the University of Chicago, reported by Stalnaker (1935). In one of the tests, five different versions of a stanza are presented, and the students are asked to mark one as the best and each of the others as poor for one, and only one, of the following reasons: too sentimental, lack of imagery, faulty rhythm, inappropriate imagery, and poor diction. The version which was most frequently (25 per cent) chosen as best by 500 scholarship applicants but which was considered by the judges to be poor because of inappropriate imagery and poor diction was:

Steeped in dust sleeps a pallid lady:
Thin and tall and blond was she;
She'd the palest hair and the bluest eyes
That ever were seen in this hot country
But loveliness dies and blue eyes slumber,
And the tomb is dark as dark can be;
And when I decay, who will remember
This lady of the West Country?

The selection which was chosen next most frequently (24 per cent) as the best but which the judges considered to be poor because of sentimentality was.

Here slumbers a wonderful mother
Bereft and sad her children three
They sadly mourn for their wonderful mother
Who taught them all to pray at her knee.
But mothers, alas, are a gift to heaven
We all must make, whoe'er we be;
And now she lives enshrined forever
In the heart of hearts of her children three.

The selection considered to be the best by the judges was selected by 23 per cent of the students, but 30 per cent of them felt that it lacked imagery. It read:

Here lies a most beautiful lady:
Light of step and heart was she;
I think she was the most beautiful lady
That ever was in the West Country.
But beauty vanishes; beauty passes;
How rare — rare it be!

Five per cent of the students got perfect scores and twenty per cent, zero. The mean was 1.7, which is only slightly above chance success.

In another type of test described by Stalnaker (1935), students were asked to match seven prose passages with seven short descriptions of the prose passages. Two of the descriptions are given below together with their passages:

Descriptions

Humility and acceptance of the orthodox have never been his weakness. Heterodoxy has always been a quality of the vigorous intellect, and his mind is electrically charged. Thus, when with nippiness and assurance he tells us what he thinks, what relationships he has discovered between the old and the new, we are ready to accept his precise and confident views. He writes with simplicity and clarity, with the firmness of tone that one expects from one whose spirit is firm and unyielding.

He has curiously mingled simplicity and gorgeousness all his own. His delight in expression communicates to the reader delight in what he expresses. He conceives them only by the desire to render his thought clear and concrete. No one can, like him, pile up splendor of description, exotic richness of phraseology, color, tones instinct with music, and then turn in an instant to a sober, solemn, stately simplicity, direct and appealing like the call of a herald. He makes life a procession to the grave but crowns it with garlands.

Passages

When I was a child and was told that our dog and our parrot, with whom I was on intimate terms, were not creatures like

> myself, but were brutal whilst I was reasonable, I not only did not believe it, but quite consciously and intellectually formed the opinion that the distinction was false; so that afterwards, when Darwin's views were first unfolded to me, I promptly said that I had found out all that for myself before I was ten years old; and I am far from sure that my youthful arrogance was not justified; for this sense of the kinship of all forms of life is all that is needed to make Evolution not only a conceivable theory when he preached to the fishes, and St. Francis when he called the birds his little brother. Our vanity . . . had led us to insist on God offering us special terms by placing us apart from and above all the rest of his creatures.

> But as when the sun approaches towards the gates of the morning, he first opens a little eye of heaven, and sends away the spirits of darkness, and gives light to a cock, and calls up the lark to matins, and by and by gilds the fringes of a cloud, and peeps over the eastern hills, thrusting out his golden horns, like those which decked the brows of Moses when he was forced to wear a veil, because himself had seen the face of God; and still while a man tells the story the Sun gets up higher, till he shows a fair face and a full light; and then he shines one whole day, under a cloud often, and sometimes weeping great and little showers, and sets quickly; so is man's reason and his life.

Only 4 per cent of 500 students of English who were scholarship applicants received perfect scores and 21 per cent received a score of zero. The mean score was 2.3 correctly matched out of 7.

In an attempt to substitute actual discrimination for verbal vagaries, Cannon (1937) selected two short passages of approximately 150 words each, from each of the following authors: Joseph Addison, Ernest Hemingway, Francis Bacon, Charles Lamb, R. L. Stevenson, H. L. Mencken, Samuel Pepys, John Lyly, Jonathan Swift, and Lytton Strachey. Students were presented with the twenty passages typed in random order and asked to indicate the two selections by each author. (The authors' names were not given.) Among 51 college students, 37 succeeded in matching the passages from Bacon, 28 from Pepys, 26 from Lamb, and 26 from Lyly. Only 7 students matched the two paragraphs from Stevenson, 13 from Strachey, 16 from Swift, and 16 from Mencken. Stevenson's work was more frequently matched with Hemingway's than with his own. The average number of correct matches was 4.09.

Such discrimination-of-style tests seem to have been applied principally in college and to have proved extremely difficult even there. The technique is simple enough, however, and the emphasis on first-hand discrimination is good both in testing and in training.

B. Tests of Appreciation

In the measurement of responses to oral presentations there appear to be two somewhat independent aspects, the sound and the meaning. The sound stimuli include various consonant, vowel, pitch, timbre, and rhythm combinations. Memories of sound stimuli may also be important in individual silent reading. The meaning of a selection involves the recognition of simple words and literary references, and also judgments of their appropriateness in a total pattern. A few samples of appreciation tests will be given.

1. Sounds. The problem of determining the relative pleasantness of speech sounds was approached by Roblee and Washburn (1912), who read aloud a list of nonsense syllables which consisted of one vowel followed by one consonant. The listeners rated each syllable on a seven-point scale of pleasantness. The average ratings showed the following order of preference for vowels:

1. *a* as in *father*
2. *e* as in *get*
3. *o* as in *wrote*
4. *a* as in *fate*
5. *i* as in *write*
   a as in *hat*
6. *oo* as in *boot*
   i as in *hid*
7. *o* as in *hot*
   ee as in *feet*
   aw as in *bawl*
8. *io* as in *boil*
9. *u* as in *mud*

The differences between the best and worst were so small and the range for each vowel so large that there seemed to be no marked agreement. Consonants showed slightly greater average differences in the following order of pleasantness: l, m, n, v, th, s, z, p, d, b, zh, t, sh, k, and g. Givler (1915) found similar results from experiments in which vowels and consonants were combined in various ways. He reported that the explosive consonants and the shorter vowels were judged to be less agreeable than the others. No standard test using such judgments seems to have been made.

Downey (1927) asked college students to record their reactions to single words. In evaluating the responses she used the following classification similar to that which Bullough (1910) used with colors (A few examples from her observers are also given.) :

1. Objective responses to the sound or appearances of a word: *Murmur* – the sound is the meaning
2. Associative responses: *Drowsy* – sleepy, or *Lily* – visualization of one fall lily in a bare space
3. Physical responses within the subject: *Pendulous* – feeling of being suspended
4. Symbolic, an arbitrary association: *Melancholy* – a green and purple word
5. Personalized: The word is treated as a person: *Twilight* – word looks half asleep

She found no marked tendency, as Bullough had, for a person to use mainly one type of response. The three subjects who made the largest number of symbolized and personalized responses were those much interested in literature and visual art, but the most "literary" person made many objective responses. No clear relation between type of response and literary ability was demonstrated, but the test needs to be refined and enlarged to give it reliability and internal consistency. This method may possibly be made to appraise important types of responses better than other methods.

2. PROSE AND POETRY. A number of tests have appeared which ask a person to indicate relative literary merit among

a series of short samples. In none of these is any definition of literary merit provided; hence emotional appeal, content, clarity, and style are probably effective in unknown amounts. However, in view of the difficulty of analyzing these elements, the standard tests doubtless represent a marked advance over cruder methods of rating individuals' literary appreciation.

Three test forms were prepared by Carroll (1935), one for use in junior high school, another in senior high school, and a third in college. Each form consists of a fourteen-page booklet. On each page four short paragraphs of approximately 100 words each are printed. The paragraphs on each page are all concerned with the same topic. Thus, in the college test the topics are *A Man, An Interior, A Sunset, A Fire, A Tryst, Spring, A Woman, Homecoming, Wind, Literary Criticism, Twilight, Remarks to a Son, Delirium,* and *Sunrise.* The paragraphs on each page were selected to represent four levels of literary excellence on the basis of reputation of the authors and judgments of experts (see Table 71). No elaborate analyses of judgments were made to insure equally-often-noticed steps between the levels of literary excellence, but the ranking of the four samples by judges showed marked agreement in making first and last choices, and fairly marked agreement in selecting second and third choices. The student is asked to rank the four selections on each page in order of his estimate of its "literary merit." The pages are arranged in order of ease of discrimination ranging from the easiest, which was correctly ranked by approximately 60 per cent of a group, to the most difficult, which was ranked correctly by 30 per cent only.

Both split-half reliability and retest reliability for groups of 300 students were approximately .70. In scoring, 2 points were allowed for each correct ranking of the best and worst selections and 1 point each for the second and third selections. The total score for each page was therefore 6, and for the whole test, 84. The mean scores for college fresh-

TABLE 71

Samples of a Prose Appreciation Test

An Interior *

A

I went with the little maid into a gorgeously decorated bedroom, all of cream color and light blue that blended prettily. The bed was a great, wide affair of beautifully carved and ornamented wood, painted creamy white with blue and gold trimmings. There was a wonderful bureau and a dressing table to match, and in one corner ot the room a mirror that went from floor to ceiling. I had to hold my breath.

B

Lollie had never seen such a pretty room, and it made her gasp to see how pretty the furniture was, as well as how pretty the rugs were, and the curtains at the windows and the pictures on the wall, but what she really liked best was that furniture, for it looked comfortable as well as pretty, and she knew it must have cost hundreds and hundreds of dollars. She wished she could live and die in that one room, it was so pretty.

C

An air of Sabbath had descended on the room. The sun shone brightly through the window, spreading a golden lustre over the white walls; only along the north wall, where the bed stood, a half shadow lingered. . . . The table had been spread with a white cover; upon it lay the open hymn book, with the page turned down. Beside the hymn book stood a bowl of water; beside that lay a piece of white cloth. . . . Kjersti was tending the stove, piling the wood in diligently. . . . Sorine sat in the corner, crooning over a tiny bundle; out of the bundle at intervals came faint, wheezy chirrups, like the sounds that rise from a nest of young birds.

D

Major Prime had the west sitting-room. It was lined with low bookcases, full of old, old books. There was a fireplace, a winged chair, a broad couch, a big desk of dark seasoned mahogany, and over the mantel a steel engraving of Robert E. Lee. The low windows at the back looked out upon the wooded green of the ascending hill; at the front was a porch which gave a view of the valley.

* Key: First, C; second, D; third, A; fourth, B.

(Carroll, 1935, p. 3. By permission of the Educational Test Bureau, Minneapolis, Minn.)

men were 45.9, for sophomores, 47.0, for juniors, 49.8, and for seniors, 52.8. Centiles are furnished for each class in college. Grade norms in junior and senior high schools are also provided for their respective forms.

A test designed to measure poetry appreciation was published by Rigg (1937). Forty short selections by recognized poets were paired with forty other inferior selections on the basis of similarity of central thought. The task was to choose the best selection in each pair. Two equivalent forms were found to correlate .815 among 342 college students.

A multiple-choice test of literary appreciation was described by Fox (1938). The test requires that omissions in various passages of prose and poetry by famous authors be filled in by choosing one of four alternatives. It was found that graduates in literary subjects choose the words written by the author more frequently than graduates in nonliterary subjects. From introspections Fox concluded that the students of literature employed a "literary feeling" and the others tried to exercise "critical judgment."

C. Tests of Information

Both discrimination and appreciation of literary work is probably to a large extent based on memory of the literature which one has read. A widely-read person with a good memory will almost always make finer discriminations than one with equal ability who has not enjoyed the same experiences. Furthermore, although a certain literary work may not be associated with a particular individual's pleasure, still a wide knowledge of literature will certainly affect his feelings toward various masterpieces. From these considerations it follows that discrimination, appreciation, and information in the field of literature are all intimately connected. Quite a number of tests about authors, literary characters, and characteristics of literature are available. Typical of these are the Jordan and Van Wagenen (1933) Scales of Attainment in Literature for Grades VII to XII, and the Cooperative Literary Acquaintance Test of Beers and Paterson (1933) for secondary schools and colleges. Both are composed mainly of five-choice items which deal with books usually assigned in literature study courses.

The first yields a total *Literary Age* score similar to a mental age and also separate scores for evaluating the emphasis in learning, as follows:

1. Information about literature, shown on such items as:
   The poem *Miles Standish* is about: 1) a shipwreck, 2) an exploring party, 3) an Indian attack, 4) a courtship, 5) scattering a settlement.
2. Information about authors:
   The Lady of the Lake was written by: 1) Shakespeare, 2) Scott, 3) Dickens, 4) Cooper, 5) Stevenson.
3. Outcomes of situations:
   In *Snowbound,* the family spent their evening in: 1) listening to stories, 2) dancing, 3) reading books, 4) playing cards, 5) reading newspapers.
4. General impressions and characters:
   Holmes' poem *How the Old Horse Won the Bet* is: 1) joyful, 2) inspiring, 3) sad, 4) mysterious, 5) humorous.

The Co-operative Literary Acquaintance Test yields a total score and also separate scores for: (*a*) Pre-Renaissance and Foreign, (*b*) English and American from 1500 to 1900, and (*c*) modern English and American. It consists of twelve pages which can usually be answered in about forty minutes.

Tests of this sort are by far the most widely used in the field of literature. Research is needed to show the interrelationships of abilities to discriminate, appreciate, remember, and compose literary work.

D. Appraisals of Compositions

Excellence of composition is the concern of a host of teachers of literature in elementary schools, high schools, and colleges. Compositions are usually rated by judges for various aspects. Upon the language aspects, such as punctuation, grammar, and diction, judges are often able to agree fairly well, but perfect agreement upon aspects of style or taste is rare even among persons with very similar training. At the earlier ages correct language usage is often stressed, perhaps to the detriment of poetical or literary development. At later ages literary excellence is usually stressed. Many books have been written upon the nature of literary excellence, but no standardized schema of appraisal except those noted in chapter 9 have come to hand.

E. Analysis of Motives

The analysis of motives for artistic activity is of great importance to those interested in human adjustment, both normal and abnormal. More than fifty somewhat technical articles have been written by psychiatrists in the last few years which show relationships between psychotic behavior and literary expression. Typical of these is the report of Levey (1938), who gave a detailed account of a patient's production of poetry during a period of psychoanalytic treatment. The poetry reproduced symbolically the childhood defense against acute anxiety and illustrated the various defensive solutions which had been attempted.

A fairly large number of analytical studies have also been made of well-known authors and their works. Illustrative of this approach is Duff's (1937) review of significant events in the life of Jonathan Swift. Emphasis is placed, according to psychoanalytic theory, upon pregenital fixations which seem to be manifested in his type of fantasy, and upon the peculiar forms of insight into himself which appear in his writings.

II. MEASUREMENT IN MUSIC

Standard measures in the field of music, as in other arts, include fairly precise tests of the simpler discriminations and skills, but only rough appraisals of more complex appreciations and of creative work. The last thirty years have witnessed a steady expansion of measuring techniques and their uses. Standard measures may be conveniently classified among the following: (*a*) Discrimination and memory, (*b*) Appreciation or satisfaction, (*c*) Creative composition, and (*d*) Accomplishment. Most of the published research seems to be concerned with the first and second groups. No standard appraisals of composition ability have been discovered, but samples of each of the other types will be described.

A. Discrimination

Probably the most widely used tests of musical discrimination are those of Seashore (1919). By means of standard phonograph records one may secure scores of accuracy of pitch, loudness, rhythm, consonance, and tonal sequence discrimination. A recent revision of the Seashore tests (1939) has three levels of difficulty or series. Series A is for unselected groups, Series B for musical groups, and Series C is a still more refined instrument for individual testing. Three 12-inch phonograph records are provided with a complete test on each side of each record. Separate scores are secured by standard testing procedures for:

1. *Pitch:* discrimination of small differences between the frequencies of vibration of two tones with intensities and duration held constant
2. *Intensity:* ability to judge which of two tones are the louder, with pitch and duration constant
3. *Time:* ability to indicate which of two short periods are the longer. The limits of the periods are marked by a pure tone of short duration
4. *Timbre:* discrimination between two complex tones which have the same total energy, but which differ in the application of the energy
5. *Tonal Memory:* ability to indicate which note in a short melody has been changed in its repetition. One note only is changed by a whole tone
6. *Rhythm:* the ability to discriminate short rhythmic patterns

The scores of fifth-grade, eighth-grade, and college students, nearly 1,000 of each, were tabulated by Seashore, and then converted into centile ranks. The median scores showed marked differences although the groups overlapped almost 100 per cent. For instance, in Pitch the median fifth-grade pupil had about 67 per cent of items correct; the eighth-grade pupil, 78 per cent; and the college student, 81 per cent. The eighth-grade scores were thought to be typical of average adults. The reasons for differences among these groups are not clear. It may be that the younger groups make poorer scores because of inability to understand the directions or unwillingness to pay attention. On the other hand, it may be that acuity of discrimination improves by maturation during adolescence.

The effect of repetition of the Seashore measures of musical talent has been reported by Farnsworth (1931) to be very small under the standard test conditions. C. E. Seashore (1919) and Brennan (1926) have also reported that special musical training has little effect upon test scores. On the other hand, Whipple (1903) and R. H. Seashore (1926) reported marked improvements through training in pitch discrimination and rhythm, respectively. More research is needed to show the limits of individual improvement, particularly among those who at present make low scores because of poor interest and lack of opportunity to hear music which has been accurately reproduced.

A number of similar tests have been prepared. Kwalwasser and Dykema (1930) published ten phonographic tests, the first seven of which are much like the Seashore tests. The last three involve judgments of pleasantness and association of images with tonal and rhythmic sequences. Ortmann (1929) devised tests of discrimination of pitch, intensity, and fusion, and of memory for pitch, rhythm, melody, and harmony. Tests of absolute pitch have been described, but not yet standardized. Bachern (1937) tested 103 persons for absolute pitch by asking them to identify notes struck on a piano. Seven of them were correct for every piano note, and for other musical instruments as well. Forty-four made only small errors, or errors of exactly one octave. He proposed a theory which makes absolute pitch dependent in part upon "tone chroma" within the octave.

B. Appreciation of Musical Stimuli

In an analysis of musical appreciation, Moore (1914) studied reactions to pairs of tones. He pointed out that the pleasantness of a musical interval depends to a marked degree upon use. His experiments seemed to show that the keenest pleasure was experienced in chords which one has recently learned to like, and he interpreted this fact as attributable to a satisfaction in the accomplishing of a somewhat difficult unification. Chords which were easily unified were regarded as trite or dull, and those which were too difficult to synthesize were experienced as noisy and unpleasant. He pointed out that, historically, *a third* had at first been considered too harsh for any musical use. Later the same interval had been only rarely used in the middle portions of sequences. Then, after a transition period, it was used as a parallel interval, and lastly as a final chord of sequence. Moore's listeners reported that habituation to thirds brought about a rapid loss of pleasant affect. Major and minor sevenths became more pleasant with repetition, and fifths remained fairly constant.

A more discriminating analysis was made by Guernsey (1928), who asked groups with various amounts of musical training to judge stimuli for:

1. Fusion: one or two notes heard
2. Smoothness
3. Pleasantness

Tuning forks with resonators were used to give twelve intervals of the tempered scale. In all her groups she found a marked agreement between reports of fusion and smoothness. The pleasantness of intervals varied with the amount of musical training, although all groups, excluding the extreme individuals, had the same marked likes and dislikes. The less trained listeners preferred simpler intervals than the more trained observers. The more trained listeners also made more accurate analyses of what they heard and seemed to be more sensitive to overtones. She also reported that the order of presentation affected the analysis slightly.

Following the idea that intervals may have characteristic responses of various sorts, Edmunds and Smith (1923) asked subjects to match verbal phrases on a printed sheet with tuning-fork tonal pairs. They report fairly close agreement in matching this phrase, "rich, like whipped cream or ripe cantaloup," to a particular chord.

Myers (1927) asked fifteen persons to listen to phonograph records and to report their responses. He classified these, following Bullough's (1910) color experiment, into:

1. Objective reports of the sounds and methods of playing, such as: "I noticed the second horn was too loud," or "I always want to know how the conductor is getting his effects."
2. Intra-subjective reports of one's own physiological responses, such as: "I felt warm in the ear," or "I had a lazy feeling."
3. Associative reports of scenes among which the most frequent were: dancing savages, lovers, fairies, goblins, a stage and religious scenes.
4. Reports of the character of the music; types of activities and imputed moods which would ordinarily be aroused. The piece may be given a personality with qualities, such as "morbid, jovial, insincere, dainty, mystic, reckless, playful, etc."

Myers regarded the reports on the character of the music as the highest type of aesthetic response. He believed that the objective responses were critical; the intra-subjective responses, crude; and the associative responses often irrelevant.

In another experiment, Schoen and Gatewood (1928) asked 32 college girls to check one of a list of adjectives to show how each of ten phonograph reproductions made them feel. They reported a marked uniformity of feelings among the groups, and also a marked correlation between first and second trials on a number of the selections. Rigg (1937) found that 20 musical passages were classified fairly uniformly by 100 students as sad or joyful, but that hopefulness and love were not uniformly interpreted from any passages.

None of these investigations have led to standardized tests, but they show quite clearly that musical discriminations and associations have marked group trends. Schoen (1928), however, published a test of musical "feeling and understanding to be played on the piano." Four musical phrases were selected from standard compositions and each was provided with four terminal phrases. The student was asked to indicate his preference for terminal phrases, after a group of four had been played. Familiarity with the original melodies may influence the results, and values assigned to the terminal phrases have not been widely agreed upon. The method, however, is a sound one.

In Kwalwasser's (1926) tests persons are asked to indicate which of two sequences they prefer. The Melodic test consists of 35 pairs of single-tone sequences, and the Harmonic test consists of 35 pairs of three-chord sequences. The scores indicate one's agreement with a group of musical experts.

One other rather rough measure of appreciation should be mentioned. Trabue (1923) reported a standardized presentation of 16 phonograph records of different composition played by an orchestra. Four records were played in a definite order; then each person was asked to rate them for merit of composition. Next, four other records were played and rated. This was continued until the series was finished. Scores were determined by the comparing of ratings with an order assigned by a group of expert musicians. The discrimination became harder as the series progressed.

C. Accomplishment

Appraisals of musical accomplishment readily fall into two classes: tests of information and of rendition.

1. Information. Several information tests have been standardized. Those of Kwalwasser are typical of the best. The Kwalwasser-Ruch (1924) Test of Musical Accomplishment was designed to measure ability to read musical nota-

tion from the fourth to the twelfth grade. Ten separate tests were provided to indicate:

1. Recognition of names of musical symbols
2. Recognition of names of notes in a scale
3. Detection of pitch errors in a familiar melody
4. Detection of time errors in a familiar melody
5. Recognition of pitch names
6. Knowledge of key signatures
7. Knowledge of time signatures
8. Knowledge of note values
9. Knowledge of rest values
10. Recognition of familiar melodies from notation

The items were arranged within each test according to order of difficulty and the mean scores for 4,000 children were reported. Most of the tests showed steady progress throughout the grades, but Kwalwasser found some evidence that the last two tests showed little improvement beyond the seventh grade. Girls did slightly better than boys, but the average child was found to be seriously lacking in most of the simplest fundamentals.

Another test by Kwalwasser (1927) measures verbal information about composers, compositions, and instruments. About 200 true-false and completion items, such as those shown in Table 72, were designed for adults who had taken courses in musical appreciation.

2. Rendition. Ability to play or sing artistically has been the subject of a number of analytical studies. Seventy-four reports of appraisals of the sound of spoken prose or poetry were reviewed by Metcalf (1938). Early in the century Scripture (1902) used a vibrating membrane to transfer the vibration of the voice to a kymograph record. Shepard (1913) enlarged this apparatus to include separate records from nose and mouth for vibration and the passage of air. Seashore (1927) described an instrument which recorded photographically the pitch and intensity of vocalization over long periods of time. These instruments have been used both for the analysis of the time and intensity

TABLE 72

Musical Vocabulary Test

TEST 9

Directions: The statements found below are either true or false. Read each statement carefully and then underline *true*, if the statement is correct, or *false*, if it is incorrect. The sample is marked as it should be:

Sample: Allemañde is a German dance . . . . True False

---

1. The fantasia is a composition of strict form . . . . True False
2. The polonaise is a Polish dance . . . . True False
3. The oratorio is a scriptural or epic story set to music . . . . True False
4. The coda is found in the middle section of musical compositions True False
5. Beethoven substituted the rondo for the scherzo form in the sonata . . . . True False
6. An etude is a study or lesson . . . . True False
7. A waltz is written in 2–4 time . . . . True False
8. A minuet is written in 3–4 time . . . . True False
9. A scherzo is written in 2–4 time . . . . True False
10. A note represents a time value as well as pitch . . . . True False
11. A phrase is longer than a period . . . . True False
12. Polyphonic means literally "many voiced" . . . . True False

(Kwallwasser, 1927, p. 97. By permission of the Bureau of Educational Research, University of Iowa.)

pattern of a passage and for the comparison of individual renditions. No simple methods of appraising either a passage or its rendition have appeared as yet. Birkhoff (1933) proposed an elaborate formula for comparing the aesthetic value of passages, in which *orderliness* is divided by *complexity*. Orderliness referred to alteration, assonance, rhyme, and musical vowels, and complexity was the total number of speech sounds plus the number of word junctures which did not admit of liaison. Beebe-Canter and Pratt (1937) reported a correlation of .75 between Birkhoff's aesthetic measure and students' preferences for nonsense passages. When meaningful passages were used, the correlation dropped to .08.

Studies of the artistic rendition of a passage are illustrated by Schramm (1935). He reported that short rhythmical intervals in poetry may deviate from equality by as much as 24 per cent without impairing the rhythmical effectiveness, and listed a large number of factors which contribute to rhythms, such as repetitive character of melodies, pitch of related or rhymed syllables, and precision and slurring of enunciation, of syllables, and of phrases. The best artists were considered to be those who knew the exact metrical rhythms but who deviated from them in a characteristic manner which usually defied any simple formulation.

Standard tests of sight-singing ability have been devised by Gaw (1928), Salisbury and Smith (1929), and others. The scores are a summation of errors in pitch and time, as well as omissions, hesitations, repetitions, and extra notes. Experienced judges are needed for careful examining, but the method yields fairly reliable scores.

Rendition of piano and violin music has been extensively studied and reported in a series of monographs from Iowa University and also in Seashore's (1938) *Psychology of Music*. Photographic records of pitch, intensity, and timing have been made with great accuracy, and have led to minute comparisons of the performances of great artists as well as others.

D. Correlational Analyses

The reliabilities of tests of musical discrimination and preference have been extensively reported by Farnsworth (1931), who summarized eighty-eight published studies. He concluded that Seashore's test of pitch and total memory appeared to have sufficient reliability for diagnostic purposes. The Kwalwasser tests and the remainder of the 1919 Seashore tests "should be employed with extreme caution." The following retest reliabilities were reported for the Seashore tests: pitch, .75; intensity, .66; time, .51; consonance, .65; rhythm, .47; and memory, .83. The Kwalwasser tests of melody and harmonic preferences had reliabilities of approximately .42 and .21, respectively. The entire Seashore battery had a retest reliability of approximately .88 for large single age groups.

The intercorrelation of the various tests showed similar results for groups of elementary-school and high-school pupils and adults in several studies. The median intercorrelations of the Seashore tests were approximately .48 for college groups and .25 for lower grades. Tests of Tonal Memory and Pitch showed somewhat higher correlations, as did tests of Rhythm and Time, and Rhythm and Tonal Memory. Nearly zero correlations were reported between tests of Intensity and Consonance. These correlations are, of course, affected by the different reliabilities of the separate tests, and they would be higher if the tests were more reliable.

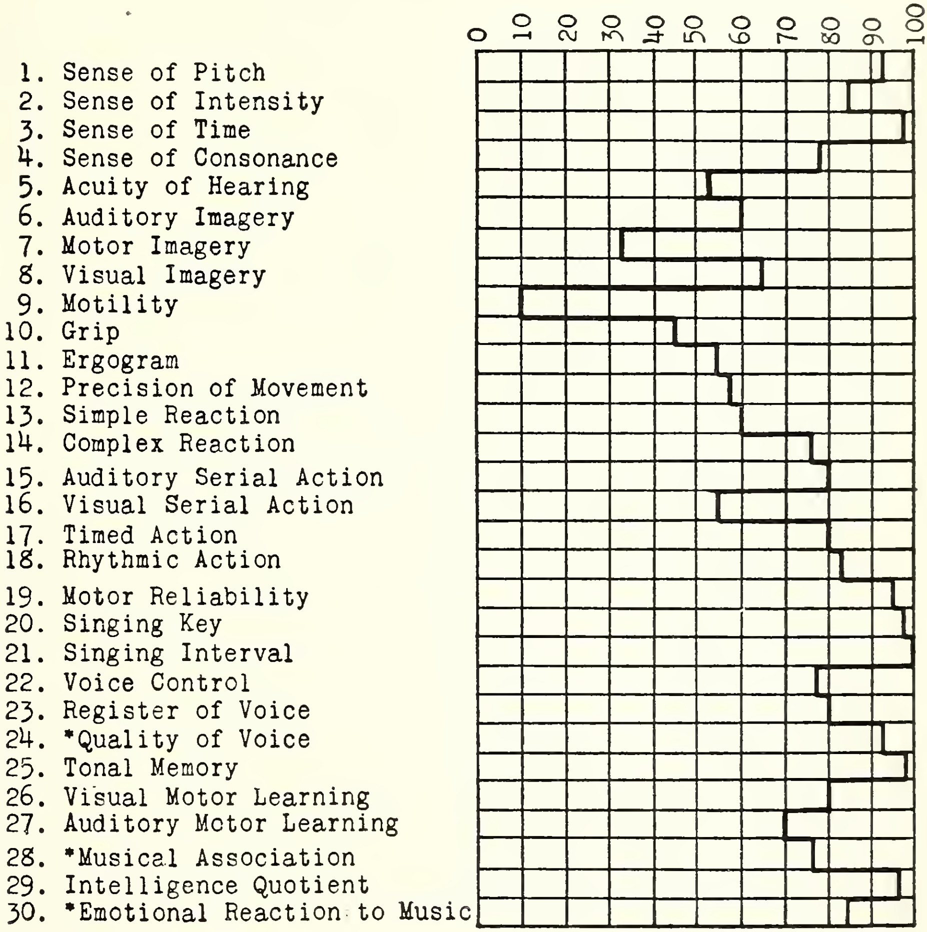
The Kwalwasser tests of preferences for Melody and Harmony correlated with each other .40 in an eighth-grade group and .29 in a fifth-grade group. Of the Seashore tests, Tonal Memory correlated with the Kwalwasser tests a little better than the rest, but all the intercorrelations were rather low; the median was .16 for the eighth grade and .17 for the fifth grade. The fact that adult and college groups usually show higher reliabilities and intercorrelations than younger groups is thought to be best explained by their better ability to pay attention and to follow directions.

The Ortmann Musical Discrimination Tests are reported by Petran (1937) to have reliabilities similar to the Seashore tests. Among a sample of 500 students, the total scores had a retest reliability of .80. The highest retest reliability of a single test was .86 for Pitch Memory and the lowest was .30 for Melody Memory. The intercorrelations of the separate tests ranged from .11 to .47.

The correlations between tests of intelligence and of musical discrimination have usually been found to be low and positive. Farnsworth (1931), summarizing sixteen such studies, reported coefficients between the separate Seashore tests and group mental test ranging from .45 to −.08 among high school and college groups. The median, which seems to be approximately .10, would doubtless be higher if feebleminded and brilliant students were included.

E. Applications

In Seashore's pioneer work (1919) an elaborate series of thirty tests and ratings were combined into profiles such as that shown in Figure 63. The first five tests were measures of fine discrimination; the next three involved self-estimates of clearness of memories of sensations. Tests 9 to 19 inclusive were measures of motor reactions. Items 20 to 24 appraised vocal rendition. Items 25 to 28 were



* The items marked with an asterisk represent estimates in the absence of norms.

FIGURE 63. **Musical Talent Chart of Theodora**

(Seashore, 1919, p. 19. By permission of Silver, Burdett and Co.)

measures of simple and complex learning of note associations, and Test 29, a standard IQ from a Binet test. Item 30 was a rating of emotional reaction to music. The chart is scaled in deciles of an age group, and the profile was made from records of a 10-year-old girl with great musical talent. Since 1919 nearly all the tests have been made more accurate and some attention has been given to improving ratings. In addition, tests have been devised for preferences among musical compositions.

Elaborate profiles have not been used very widely because of the difficulty of constructing them, but nearly everyone who has used them cautiously has reported that they yield information which is valuable to both instructor and pupil. Most of the reports of musical measurements have dealt with the correlation of a few discrimination tests with grades or ratings in various sorts of musical appreciation or skills. A typical report is that of Petran (1937), who reported that total scores on the Ortmann tests of tonal discrimination and memory correlated .57 with average grades in the Peabody Conservatory of Music. The test included measures of discrimination of pitch, rhythm, and fusion; and measures of memory for pitch, rhythm, melody, and harmony. The highest correlation of a single test with conservatory grades was .45 for *chord memory,* and the lowest was .25 for *rhythm memory*. Larson (1938) reported enthusiastically on a testing program in the Rochester, New York, public schools and the Eastman School of Music. The Seashore test scores correlated .59 ± .04 with grades in the first course of musical theory in the School of Music.

There have also appeared some critics of the uses and alleged claims made for the Seashore tests. Moos (1930) believed that they are not measures of musical talent because they stress sensory patterns rather than complex compositions and because they have been found to give largely negative results. Pratt (1931) felt strongly that they utterly failed to get at the kernel of musical talent, although admitting that they were good tests of sensory discrimination. He pointed out that the usually low correlations between discrimination tests and ratings of musical ability indicated a nearly chance relation between test scores and achievement. Mursell (1937) quoted a number of studies which showed low correlations between Seashore tests and appraisal of musical ability, and concluded that the tests had been shown to be invalid.

The writer feels that although undoubtedly there have been some exaggerated statements about the usefulness of the tests of musical discrimination, still they are of value in appraising skills which may be of importance in predicting some aspects of musical success. It seems probable that particular tests of musical discrimination will predict success in particular types of musical skill better than in musical ability of all types. There is need for analytical appraisals of musical memory and complex skills and for careful studies of their growth.

Although musical discrimination and verbal intelligence tests show very small correlations, both have been found to correlate significantly with success in schools of music. Highsmith (1929) found that grades in a college of music correlated .423 with scores in the Terman Group Intelligence Test, but only .312 with the Seashore tests. Stanton (1929) found it advisable to add a college intelligence test to the Seashore Tests of Musical Talent for purposes of selecting students for the Eastman School of Music.

Two reports which deal with scores of various racial and age groups are included here. The results were obtained from such small groups that they are in need of supplementation. Scores of southern negroes on the Seashore Tests of Musical Talent were discussed by Bean (1936). He reported that the negroes in high school and college were found to be equal to the whites only in tests of rhythm. In all other tests they were definitely inferior. On the other hand, Lenoire (1925) reported 200 negro children in the fifth grade had mean scores on the Seashore tests which were definitely superior to a similar group of white children in rhythm and tonal memory, and not inferior in any of the other tests.

15

ACADEMIC AND VOCATIONAL INTERESTS

I. DEFINITIONS

Constructive research in motivation, as in all other fields, requires very clear definitions of that which is to be measured. Motivation, however, is notoriously hard to define, partly because of the complexity of the behavior involved and partly because of widespread unanalytical ways of thinking. A large number of words are used to describe motives. Some of the most common among them are the following: attitude, want, interest, prejudice, aversion, opinion, preference, belief, like, and feeling. All of these terms, like the word *motive,* may be used to refer to reactions toward or away from some goal, or memories of a goal. They also refer to potential reactions. We often hear such expressions as, Frank has an aversion for mathematics, or Mary has an interest in gardening. Actually, there is no *measurable* thing called an interest or motive but only a series of acts toward or away from some goal. Words, such as interest or aversion, are valuable for rough classifications of goal-seeking activities. They will not be technically distinguished from one another here, but they will be distinguished from structures, ability, and modes of adjustment. *Structures* refer to physical phenomena. *Abilities* refer to patterns of behavior which have been developed. *Motives* are defined as tendencies to use abilities in attaining a goal, and *Modes of adjustment* are ways of approaching a goal, such as energetic, courageous, cheerful, or despondent. Abilities have been discussed in Chapters 8 to 12. Chapters 13 and 14 have described appraisals of preferences for artistic activities or objects. This chapter discusses academic and vocational motives and Chapter 16 describes studies of attitudes toward principles, institutions, and races. Modes of adjustment are discussed in Chapters 17 to 19.

Thinking in this field can often be made more accurate by the use of operational definitions. Operational definitions always refer to a particular situation. Thus, in physics a force of one dyne is defined as that which moves a mass of one gram a distance of one centimeter in a particular region of the earth's surface. All other forces can be measured by comparing them with one dyne. Work is defined as the total force or energy exerted by a mechanism during a period of time, such as an hour.

In the appraisal of human behavior, a motive may similarly be defined as that which tends to move a person a certain distance in a particular direction when he is in a particular situation. In order to measure a motive one must give quantitative values to its strength and direction. Motives can be measured by placing a person in a standardized situation and securing some indications of the amount of work done in attempting to reach various goals. This is the usual procedure in experimental laboratories. Experimenters have found that motivation, when defined as energy expended in attempting to reach a goal, is dependent upon two somewhat different factors called drive and incentive. *Drive* is defined as an internal tension or irritation, and *incentive* as an external goal, which relieves the tension when achieved. For instance, a drive would be a tension or dryness in the throat and other tissues, and an incentive would be a glass of water, or the memory of a glass of water. Strength of motivation in this situation could be measured by (*a*) the person's total expenditure of energy, or, more roughly, (*b*) the amount of punishment he endured to secure water, or (*c*) the time during which he continued to search for water, or (*d*) a self-rating of thirstiness.

The most careful worker tries to evaluate separately the expenditure of energy which is due to drive and that which is due to incentive. This differentiation is desirable because two persons might show the same amount of goal-seeking activity for different reasons. One might have been deprived of water for 48 hours and therefore have a very intense drive; the other might have been deprived of fluid for only 3 hours but have an intense liking for the particular fluid that was offered. In the appraisal of behavior it is desirable to secure scores which represent known quantities of particular patterns of behavior, not an unknown mixture of several patterns. Logically, there are two procedures which may be used: one measures differences between individuals in strength of drive when incentive is held constant, and the other evaluates the strength of various incentives when drive is held constant. Both procedures are used in laboratories, but usually in normal social situations neither drives nor incentives can be clearly known or well controlled. The appraisals of motives described below are typical of partly controlled situations, which usually do not make possible a distinction between drives and incentives. They yield scores which depend upon unknown combinations of both. This is a weakness which should be kept in mind when interpreting scores, and if possible eliminated in future measuring instruments. The present scales for measuring motivation do have considerable value both in pointing the way toward needed refinements and in predicting a person's probable development. They yield scores which indicate intensity of goal-seeking activities in particular fields.

II. TYPICAL STUDIES

Since motives or energy expenditure cannot be directly observed, it is necessary to regard the activities of an individual and then to infer his direction and amount of energy expenditure. Such inferences are, of course, subject to errors of various sorts, so that some writers have despaired of ever securing accurate appraisals. The results of actual studies, however, show that the most careful appraisals of attitudes and interests are judged to be nearly as accurate as the best appraisals of mental abilities. Three types of appraisals are common: case histories, studies of reasons for choices, and inventories of preferred activities. All have their advantages. A case history or a careful autobiography has the advantage of showing temporal sequences, but it has the disadvantage of not allowing accurate numerical comparison among persons. A series of case histories will first be presented. These will be followed by accounts of studies which deal with the reasons for choices and with inventories of pleasure in particular activities.

A. Case Histories

The following case histories illustrate the complexity of factors which contribute to vocational choices (Fryer, 1931).

SUBJECT 33: Male; Education, High School Graduate; Age, about 50 Years

As I think back, it seems to me that I have just gone from one relationship into another relationship, each growing naturally out of the previous situation. I left college when I was preparing to be a surgeon. Because of family objections to a choice of that vocation, I went into business without any particular choice of vocation. I began at about 20 years of age and continued in business until I was about 32, making one change during that period. This change did not involve a change of vocation which was that of salesman; finally, I became an expert in style patterns.

At the age of 32 a secretary of the Y.M.C.A. was foolish enough to see some qualifications in me for the religious work of the

Association and extended to me a call. The determining factor in that change was the facing up to all that would be involved in refusing it, and the consequent conviction of moral and spiritual cowardice. This was a period of real crisis, and for two months and a half, I studied the situation and wrestled with the problem. The final decision was reached on as clear a venture of faith as anything I have ever undertaken, and speaking very honestly, I feel that the experiences of the past 18 years have shown the wisdom of that venture in faith. I know that it was a great personal good to me to change from the interest in things and selling things to the interest in men and the putting across of ideas and ideals. The mental reorganization that was necessary, compelling me to go through about five years of hard study, gave me a new mental freedom and developed within me a capacity of which I was not aware. The settling of questions on the basis of deep inner convictions has almost become a habit with me and that is why it is now difficult to think back to the period when very many choices were dictated by chance and opportunism.

Subject 33 gives an account of a life that began on a casual and opportunistic basis. The vocation of surgeon is relinquished and business accepted without any definite, powerful interests being involved. The branch of business likewise seems not to have been very important. The basis for the great change after thirty, into an occupation of social service, is not clearly given but the appeal appears to have been unusually powerful. The resulting success in the new work, following success in the business world, suggests that more of the same qualities are involved than at first sight appears to be the case. However, a very definite shift in interest is obviously required to bring the salesman's capacities into the service of the Y.M.C.A. secretary.

SUBJECT 34: Female; Education, A.B. Degree; Age, about 25 Years

High School, 15 to 18½ years.

Early in high school, possibly during first year, I became ambitious to become a nurse. This ambition persisted and increased in intensity of desire until I was about 20 years of age and in the second year of university. Reasons for change: (1) marked opposition to nursing profession by my parents; (2) unexpected opportunity to complete university course; (3) gave up idea of becoming public health nurse because I flunked chemistry, a requisite course at this time.

University Life, 19 to 23 years.

Early in college my aspirations to become a nurse waned, and I aspired to higher goals. I next went through a period of intense desire to become a writer, colored partly by the fact that I had for instructor in English, a poet who took a personal interest in my ambitions. This ambition I still harbor, but without any very definite ideas or plans for the future.

At about 21 or 22 years of age, I decided that I would like above all else to become a physician. This was impossible because of economic pressure so I went into sociology for the last two years of college, specializing in hygiene and medical social work. I always held very definitely to the idea of doing medical social work, and became part-time worker in a clinic during the last year of college. At this time (23 years) I was in the neuropsychiatry department of the clinic, and became definitely interested in that particular field. This interest still persists and I still have faint hopes of studying medicine and specializing in psychiatry. It was because of this interest that I accepted my present position, so that I could be in close contact with psychiatric work.

I have always fostered a desire to become a dancer to such a degree that it has become an avocation with me. However, I feel it might have as well become a vocation, were it not for my family prejudices and objections early in my life, which prevented my ever taking dancing lessons.

Vocational interests mentioned by Subject 34 show a definite sequence from the original interest in nursing, through medicine, medical social work, to psychology. Outside of this line of development, which has led to work related to the work of a psychiatrist, are the two esthetic interests, one in writing and a stronger one in dancing. The influence of family prejudice is shown in checking the nursing and the dancing as vocations, while the positive influence of others is shown in the writing and in the psychiatric work.

SUBJECT 36: a woman; Age, slightly over 40 Years

She had to go to work when she graduated at 13. Before this she had had aspirations to be a teacher or a nurse. She was a clerk for a number of years, after which she did war work. Further office work was followed by positions in a hospital in which she took histories, passed on admissions, etc. Difficulties with a superior finally led to her leaving this position, which she had held for several years. Since that time she has been unsuccessfully seeking work, over a period of months. In the fall she secured a part-time selling job in a department store, and was kept on after the Christmas rush. She does not like selling but she loves keeping records. She thought she would like medical social work, anything to do with hospitals or medicine.

On the Stanford-Binet she has a mental age of 16 years. On the Stenquist she made a score less than that for the 12-year-old median, but in spite of her lack of ability to do the problems she worked steadily without complaints and tried practically all of them. She had a quiet, appealing personality and claims to be very patient and hard-working. Her interest in hospitals made her eager to enter a training course for practical nursing, although this was not recommended by the psychologist. Her history is entirely opportunistic and this final attempt might have to her satisfaction been made equally as well in a variety of office positions. Her situation is such that she must undertake whatever is opened to her, and the probability appears to be that she will be faithful and industrious in anything she tries.

B. Reasons for Choices

These cases from Fryer show that educational and vocational choices were made from a fairly large number of *causes* often conflicting with one another. The goals may change very rapidly with maturation of abilities, and also with the acquisition of information and with emotional conditioning. The causes for vocational choices, therefore, deserve special study. A thorough study of this field would involve all the dynamics of behavior. A number of special studies limited to academic and vocational situations are all that can be included here. In analyses of motivation, Thorndike (1935) has suggested that unlearned likes or dislikes be distinguished from learned. He lists tentatively, without devising a scale, the sorts of activities which should be included, such as preferences for tastes, smells, bodily temperature, muscular activity and rest, courtship and love, motherhood, receiving favorable attention, successful competition, familiar surroundings and changes of scenery, and mental activity. Thorndike believed that an inventory of unlearned likes would probably not contribute as much to adult education as to the planning of early education. Items of this sort are occasionally included in studies of modes of behavior shown in Chapter 18. Since there have been few reports of reasons for academic choices, only one study will be described here.

1. Reasons for Academic Choices. A series of experiments by Gates (1930) evaluated some aspects of interests in reading. Thirty short selections were read aloud in pairs and the pupils were asked to indicate which one they preferred. Results were obtained for about 15,000 pupils in Grades I, II, and III. The per cents of pupils preferring one selection to others were averaged to give an index of general interest. Each selection was also rated on a nine-point scale for 14 qualities, and their correlations to the interest indices are shown in Table 73. Column 1 shows the correlation of raw scores and Column 2 shows partial correlations between a single factor and interest when all the 13 other factors were held constant by statistical procedures.

TABLE 73

Quality and Interest of Selections Read Aloud

Correlations of Interest and	*Raw* r	*Partial* r *
Surprise	.34	.35
Liveliness (action)	.59	.23
Animalness	.46	.18
Conversation	.52	.15
Humor	.34	.14
Plot	.51	.13
Narrativeness	.56	.08
Poeticalness	– .24	.07
Familiarity	– .09	.06
Repetition	.28	.04
Fancifulness	.31	.01
Realism	– .29	.00
Verse Form	– .23	.005
Moralness	– .16	– .15

* Thirteen other factors held constant.

(Gates, 1930, p. 84. By permission of the Macmillan Co.)

A comparison of these two columns shows a fairly marked similarity of rank order and a general reduction of coefficients toward zero in the second column. The greatest reductions appear in conversation, plot, narrativeness, and fancifulness. When partial correlations are used, the first 6 items are the only ones which appear to be positively related to interests in small children. Realism and fancifulness were both of no significance. Moralness was the only item which proved to be negatively related to interest, and the significance was small. Gates found out that the essential satisfactions of children are only very roughly shown for groups by this study. The individual interests, which will vary considerably, could be better appraised by a scale of independent factors which would yield a profile. Gates also reported marked preferences due to interests in special projects. He states that one of the most significant outcomes of his study is the finding that approximately nine-tenths of the first graders chose reading materials which followed a fairly prolonged sequence of activities. Selections were interesting to the degree to which they supplemented a child's projects. There seemed to be little evidence that stories had intrinsic interest, but rather that values were relative to the dominating goals of the pupils.

2. Reasons for Vocational Choices. Reasons for vocational preferences may be inferred to some extent from choices of various age groups. The vocational likes of 640 boys and girls from 8 to 17 years of age in Topeka, Kansas, were reported by Lehman and Witty (1927). Each person was asked to indicate which three occupations in a list of 200 he liked best. Marked changes with age were found; thus, among girls, 5 per cent of the 8-year-olds, and 32 per cent of the 16-year-olds chose stenography. In the same group, "movie actress" was chosen by 20 per cent of the 8-year-olds and 3 per cent of the 16-year-olds. Among boys the per cents increased with age for professions and decreased for such occupations as cowboy, auto racer, and President of the United

States. Fryer (1931) reported these same tendencies in a study of 181 early adolescents. The choices tended to change from exciting, romantic, artistic, or frontier occupations to those which seem more practical. This shift of choices is probably not a true shift of likes or enjoyment, but rather a realization that normal existence requires a considerable amount of monotonous, if not unpleasant work.

A number of reports which are available show marked tendencies toward impractical choices, even among late adolescents and adults. Proctor's (1923) report is typical of many others. It shows large differences between vocational and actual opportunities. He found that 91 per cent of high school students chose "white collar jobs," whereas the census classified only 16 per cent of workers in these occupations. Furthermore, 61 per cent of high school students wished to enter professions which would probably employ only 4.4 per cent of workers. A conservative estimate indicates that four or five times as many students would like to enter professions as can be at present accommodated. Interests of girls seemed on the whole more practical than those of boys.

Another approach to the evaluation of reasons for choices is that of direct questioning. Vernon (1938) derived from interviews with 47 university women students the following list of main drives: social conformity, humanism, activity, independence, security, ease, inferiority, power, and social admiration. Activity and independence were given most frequently. Different drives were sometimes found to lead to the choice of similar careers, and similar drives may lead to the choice of different careers. Knowledge of the interaction of all the factors of a situation was essential for a complete understanding of the course of action.

A study of occupations by Anderson (1934) required 673 college men to rank 25 occupational names in order of their (*a*) contribution to social well-being in general, (*b*) contribution to one's social prestige, and (*c*) probable economic return. The numerical ranks for each occupation based on group medians are shown in Table 74. There is a marked relationship between all three values. Correlations between ranks for social contribution and social prestige made by agricultural, engineering, business, and textile students were all approximately .82; between social contribution and wages, .72; and between prestige and wages, .86. The rankings of various classes in college were very similar, and two classes three years apart also gave almost identical rankings. Anderson concludes that these occupational values are very widely and uniformly held, and that they play an important part in the selection of one's vocation.

A study by Greene (1938) of college sophomores showed that 81 per cent of 278 men and 64 per cent of 320 women had made fairly definite choices. The main reasons given for these choices are shown in Table 75. From these results it appears that men were more frequently influenced by considerations of advancement than women, and that women consider opportunity for employment and service to others as sufficient reason for choice more often than men. Persons who had chosen teaching or other professions usually gave "service to others" as the main reason for their choice. The fact that most persons had not considered "ability as tested or demonstrated" or "enjoyment of the work" of major importance in choosing a career is challenging. In the long run these two factors are probably important for vocational adjustment. It points to a need for more accurate self-appraisal.

C. Reasons for Job Satisfaction

The analysis of reasons for job satisfaction allows an interesting comparison with reasons given for vocational choices. The evaluation of job satisfaction is also important in industrial relations because it allows the supervisor to take remedial measures in particular cases and to avoid sources of friction. At least fifty reports have been published con-

TABLE 74

Occupational Values

	MEDIAN RANKING		
	Social Contribution	*Social Prestige*	*Economic Return*
Clergyman	2.9	4.8	14.5
Physician	3.2	4.7	4.9
Professor	4.6	6.6	10.9
Banker	6.1	3.1	2.9
School teacher	6.4	11.6	17.3
Manufacturer	7.6	6.7	3.0
Lawyer	7.8	6.1	5.8
Farmer	8.3	14.4	13.2
Engineer	8.4	9.4	6.4
Artist	8.6	7.0	8.9
Merchant	11.0	11.7	10.5
Factory manager	11.7	11.4	8.6
Machinist	15.6	17.2	12.9
Carpenter	15.6	18.8	15.2
Bookkeeper	15.8	15.6	17.0
Insurance agent	16.3	15.4	13.9
Salesman	17.0	15.5	14.0
Factory operative	18.3	21.2	19.8
Barber	18.9	20.1	19.9
Blacksmith	19.6	21.7	20.1
Baseball player	19.8	14.2	8.9
Soldier	20.7	21.7	23.7
Chauffeur	23.0	23.1	22.5
Man of leisure	23.8	7.3	14.4
Ditch digger	24.9	25.5	24.6

(Anderson, 1934, p. 443. By permission of the Editor, *Journal of Social Psychology*.)

TABLE 75

Reasons for Vocational Choices of College Sophomores

N = 598

	PER CENT	
Reason Given	*Men*	*Women*
1. Opportunity for employment	1	15
2. Opportunity for training	2	0
3. Initial income	3	6
4. Expectation of advancement	54	13
5. Desire to be of service	18	41
6. Pleasantness of the work	5	9
7. Ability as tested or demonstrated	5	10
8. Family tradition	9	4
9. Location, part of country	2	0
10. Health	1	2

(Greene, 1938.)

cerning work satisfaction, of which three will be discussed.

A 50-item check list of attitudes of employees toward their work and the company which employed them was constructed

TABLE 76

Attitudes of Employees

Scale Value	
10.4	I think this company treats its employees better than any other company does.
9.5	If I had to do it over again, I'd still work for this company.
9.3	They don't play favorites in this company.
8.9	A man can get ahead in this company if he tries.
8.7	I have as much confidence in the company physician as I have in my own doctor.
8.5	The company is sincere in wanting to know what its employees think about it.
7.9	A wage, incentive plan offers a just reward for the faster worker.
7.4	On the whole the company treats us about as well as we deserve.
6.3	I think a man should go to the hospital for even a scratch as it may stop blood poisoning.
5.4	I believe accidents will happen no matter what you do about them.
5.1	The workers put as much over on the company as the company puts over on them.
4.4	The company does too much welfare work.
4.1	Soldiering on the job is increasing.
3.6	I do not think applicants for employment are treated courteously.
3.2	I believe many good suggestions are killed by the bosses.
2.9	My boss gives all the breaks to his lodge and church friends.
2.5	I think the company goes outside to fill good jobs instead of promoting men who are here.
2.1	You've got to have "pull" with certain people around here to get ahead.
1.5	In the long run this company will "put it over" on you.
1.0	The pay in this company is terrible.
0.8	An honest man fails in this company.

(Uhrbrock, 1934, p. 367. By permission of the Editor, *Journal of Social Psychology*.)

by Uhrbrock (1934). He used Thurstone's scaling technique, which gave the scale values shown in Table 76. This form was distributed by superintendents to approximately 3,900 factory workers, 100 clerical workers, and 400 foremen in twelve factories. The blanks were not signed and the employees deposited them in ballot boxes without supervision.

A satisfaction score was secured by finding the median scale values on items which an employee checked. The mean scores for various groups were compared. Clerical workers had a slightly higher mean score than factory workers but lower than the foremen. Persons employed for more than six years had a slightly higher score than those employed for a shorter time and women had slightly higher scores than men. The clerical workers were also asked to complete an Otis Self-Administering Examination and a 60-item test of information about the company. Small negative correlations were found between satisfaction scores and the scores of these two tests. The overlapping of scores of the various groups was marked.

A ten-year follow-up study of 1,140 persons who had passed through New York City schools was published by Thorndike (1935). On the basis of an interview each person was rated on the following scale:

Score

7 Would rather do present work than any other job he knows of
6 Surely likes (but less than 7)
5 Halfway between 6 and 4
4 Indifference, no definite feeling for or against
3 Halfway between 4 and 2
2 Surely dislikes (but less than 1)
1 Dislikes greatly (would be in despair if he thought he must continue long)

He found that the median score was slightly higher for clerical than for mechanical or manual workers, for girls than for boys, for children of immigrants who lived in poor neighborhoods than for others, and for the ages 20 to 22 than for 18 to 20 years.

An elaborate study of job satisfaction among various persons was reported by Hoppock (1935). He used two verbal self-rating forms: one, for satisfaction in a job as a whole, given in Table 77; and another, for aspects of the job.

TABLE 77

Job Satisfaction Blank No. 5

You are asked to help in a scientific study by answering the questions in this blank. Neither your employer nor any of your associates will be allowed to see your answers. Your replies will be added to those of many other people, and only the group totals will be published. Do not put your name on the paper. Your answers will be worthless unless they are perfectly frank and truthful. If for any reason you prefer not to tell exactly how you feel about your job, please return the blank unmarked.

Choose the ONE of the following statements which best tells how well you like your job, place a check mark (√) in front of that statement:

1 ....I hate it.
2 ....I dislike it.
3 ....I don't like it.
4 ....I am indifferent to it.
5 ....I like it.
6 ....I am enthusiastic about it.
7 ....I love it.

Check one of the following to show HOW MUCH OF THE TIME you feel satisfied with your job:

8 ....All of the time.
9 ....Most of the time.
10 ....A good deal of the time.
11 ....About half of the time.
12 ....Occasionally.
13 ....Seldom.
14 ....Never.

Check the ONE of the following which best tells how you feel about changing your job:

15 ....I would quit this job at once if I could get anything else to do.
16 ....I would take almost any other job in which I could earn as much as I am earning now.
17 ....I would like to change both my job and my occupation.
18 ....I would like to exchange my present job for another job in the same line of work.
19 ....I am not eager to change my job, but I would do so if I could get a better job.
20 ....I cannot think of any job for which I would exchange mine.
21 ....I would not exchange my job for any other.

If you could have your choice of all the other jobs in the world, which would you choose? (Check one):

22 ....Your present job.
23 ....Another job in the same occupation.
24 ....A job in another occupation.

Check one of the following to show how you think you compare with other people:

25 ....No one likes his job better than I like mine.
26 ....I like my job much better than most people like theirs.
27 ....I like my job better than most people like theirs.
28 ....I like my job about as well as most people like theirs.
29 ....I dislike my job more than most people dislike theirs.
30 ....I dislike my job much more than most people dislike theirs.
31 ....No one dislikes his job more than I dislike mine.

Which gives you more satisfaction? (Check one):

32 ....Your job.
33 ....The things you do in your spare time.
34 ....Have you ever thought seriously of changing your present job?
35 ....Have you ever declined an opportunity to change your present job?
36 ....Are you feeling today a true sample of the way you usually feel about your job?

The following questions need not be answered if they would enable anyone to know that this paper is yours:

37 What is your job? (For example, Carpenter) ....
38 Age at last birthday ....
39 Sex ....
40 Date ....
41 On the line below, place five check marks to show how well satisfied you were with your last five jobs. Use a separate check mark for each job. You may place each mark anywhere on the line, either above one of the statements or between two of them. If you have had less than five jobs, use only as many check marks as you have had jobs. Draw a circle around the check mark which indicates your present job.

Completely dissatisfied	More dissatisfied than satisfied	About half and half	More satisfied than dissatisfied	Completely satisfied

Report Blank Used in Survey of Job Satisfaction, New Hope, Pa., 1933.

(Hoppock, 1935, p. 243. By permission of Harper and Bros.)

The rating on the job as a whole includes responses to four questions:

1. How well do you like your job?
2. How much of the time do you feel satisfied?
3. How do you feel about changing your job?
4. How do you think your job satisfaction compares with that of other persons?

Ratings on aspects of the job were made by means of a form which included 200 items covering satisfaction with personal contacts, kinds of work, conditions of work and hours, security, wages, opportunity for advancement, ethical aspects of the work, and social esteem.

These forms were scored by assigning the numerical values from one to seven. Seven points were allowed for items indicating greatest satisfaction and one point for the greatest dissatisfaction. The split-half reliability of the total score of each form applied to 301 employed adults was approximately .87. The two forms correlated with each other .67. This correlation between the two forms indicates that a summation of ratings of various aspects is not equivalent to ratings of satisfaction with the job as a whole. Hoppock points out that a single unsatisfactory aspect may overbalance all the rest in a general rating but not in a detailed rating.

The results of applying these two forms to 500 teachers in the Northeastern United States and 309 adults in New Hope, Pennsylvania, were analyzed. The 100 most satisfied teachers, as shown by the total scores, were compared with the 100 least satisfied. This comparison showed that the satisfied reported less emotional maladjustment, fatigue, and monotony than the dissatisfied. The satisfied rated themselves as more religious, happier in relations with supervisors and associates, more successful, and more satisfied with family reactions to their jobs. Moreover, the satisfied teachers stated that they had selected their vocation instead of drifting into it. They were 7.5 years older than the poorly satisfied ones, the means being 37 and 29.5 years. When paired for age there was no marked difference between the satisfied and dissatisfied groups in length of service, weight, illness, physical characteristics, or alleged interest in the type of work. No measure of proficiency was attempted.

The survey of 309 adults was also analyzed to show mean scores of various occupational groups. These are given in Table 78. A marked increase in satisfaction is noted from the unskilled to the professional groups. There is also a marked overlapping of group dispersions. The maximum satisfaction shown by a score of 700 was recorded in six cases, artist, art workers, clergyman, dentist, storekeeper, and veterinarian. The lowest possible score, 100, occurred in three cases, two laborers and a farmer. In discussing these results, Hoppock concludes that they seem to indicate that

TABLE 78

Job Satisfaction Indices of Occupational Groups

Occupational Classification	*Number of Cases*	*Range of Indexes*	*Mean Index*
Professional men	6	625–700	667
Artists	5	525–700	615
Railroad employees	8	500–650	584
Managers, foremen, and inspectors	8	300–675	566
Clerical workers	17	325–650	538
Miscellaneous	29	250–650	534
Proprietors of small businesses	21	250–700	524
Truck drivers	5	425–600	520
Salesmen	6	250–675	517
Workers in skilled and semiskilled trades	27	125–650	506
Occupation not stated	32	150–675	501
Clerks in stores	10	275–650	492
Factory workers	41	225–650	474
Watchmen, caretakers, and janitors	5	275–650	470
Domestic servants	8	300–625	462
Teachers	11	300–650	455
Machinists and mechanics	8	125–575	453
Laborers	35	100–650	430
Restaurant and hotel employees	7	125–625	382
Farmers	12	100–500	298

(Hoppock, 1935, p. 255. By permission of Harper and Bros.)

active, conscious dislike of their job is found in one-third or less of all employed persons.

All of these studies show a large overlapping of satisfaction scores among various occupational, age, and sex groups. The writer believes that a more analytical approach is desired which would furnish a profile to show satisfaction in several independent aspects of employment. The basic aspects which should be separately appraised are not easily agreed upon, but an inspection of several investigations brings out six main factors which would form a good working hypothesis. These may be called:

1. *Mental Balance and Ideals.* Many studies report a marked inverse relationship between job satisfaction and paranoid or neurotic tendencies. They also report a positive relationship between satisfaction and calmness, ideals of service, and a practical philosophy of life.
2. *Personal Friendships on the Job.* The most frequently mentioned single factor in satisfaction is a friendly attitude of supervisors and associates. This is particularly important among those who work under close supervision.
3. *Special Working Conditions.* These include such factors as hours of work and ventilation.
4. *Enjoyment of the Type of Work.* This involves the relation between the nature of the work and personal fitness in physical, mental, and social skills.
5. *Security.* This includes a guarantee of regular employment, unemployment compensation, and pensions.
6. *Wage.* This should be considered from the point of view of the community in which the person lives as well as his responsibility to dependents.

There are several minor factors which may, in some instances, be more important than the major factors listed. They might include:

7. *Attitude of one's own friends and family toward his work*

8. *Length of service and past recognition*
9. *General social esteem for a type of work* (This factor seems to be closely related to factors 3, 5, and 6.)
10. *Age* (Although age has been found to have a positive correlation with satisfaction, the relationship is generally very low and will probably only hold for short age ranges.)
11. *Location in a particular city or part of country*
12. *Health* (Special handicaps)
13. *Opportunities for training in service*

A comparison between these factors and the reasons given for choosing a vocation are strikingly similar. The only important differences are the emphasis in the latter upon opportunities for employment and for preliminary training.

D. Inventories of Preferences

Inventories of names of activities are available for a very large number of vocational and educational situations. The simplest of these are short check lists or blanks for indicating one's first choices. The most elaborate include about 400 items of various kinds. In some cases the blanks are to be filled in by a counselor after an interview and in others, a person is to rate himself. Both interviews and questionnaires have good and bad points. The questionnaire allows a person to check a larger number of items than could be considered in a short interview, but an interview often yields a more coherent account of developments and conflicts. The questionnaire is free from whatever effect the interviewer's personality may have, but it lends itself to intentional misrepresentation which may be reduced by a careful interviewer. An inventory often shows relative strength of preference more clearly than an interview because it allows a quantitative comparison with the ratings of large groups of persons.

Today there are available at least thirty questionnaires which are much alike. They include items from academic, recreational, and vocational fields; hence they can be used to yield interest scores in all of these fields. Typical scales by Kefauver and Hand (1937), Garretson and Symonds (1930), Wallar and Pressey (1937), and Strong (1938) will be discussed. Several inventories which consider both activities and character traits will be described in Chapter 16.

1. Guidance Tests of Kefauver and Hand. A valuable approach is seen in the Guidance Tests of Kefauver and Hand (1937) for use in Grades VII to XIV. Six tests and two inventories have been prepared which allow a direct comparison of one's information and preferences. The tests were designed to show knowledge about opportunities for self-development. They do not include knowledge of technical skills. Separate scores for knowledge of school and community facilities are available in five fields: educational, health, recreational, social-civic, and vocational. A marked degree of internal consistency was secured by selecting only those items for each test which clearly distinguished the upper from the lower third of students grouped according to total scores of preliminary forms. Vocabulary difficulties have been carefully adjusted to the seventh-grade level to avoid difficulties in understanding the questions.

Group norms are purposely omitted because the authors believed that the greatest usefulness of the form lies in appraising an individual's strong and weak points. They state that there is as yet little justification for setting up norms which might be erroneously accepted as standards. The tests, however, do lend themselves easily to normative data, and hence to surveys of school systems which would evaluate student orientation and self-planning activities.

The two inventories require students to rate their own capacities, to describe goals they have selected, and to give reasons for their selection. The first, an inventory of student plans, calls for statements of specific ambitions in the five fields already tested for information — namely, educational, health, recreational, social-civic, and vocational.

Space is also provided for reasons supporting plans for both educational and extra-curricular activities during each grade from seven to fourteen. Occupational preferences, social and civic activities, recreations and plans for establishment of a home are also called for but without a yearly schedule. The second, a Student's Self-rating Blank, requires ratings of capacity in sixteen general school activities or fields of study, and ratings of actual enjoyment of fifty topics most of which are recreational. Sixty occupations, largely non-professional, are listed, and the student is asked to indicate his probable success on a 5-point scale. Thirty athletic games are to be rated as harmful or beneficial, and, lastly, a list of forty physical defects is to be checked.

2. GARRETSON AND SYMONDS' INTEREST QUESTIONNAIRE. In order to meet a strongly expressed need for an instrument to use in measuring interests of eighth- and ninth-grade children, Garretson and Symonds (1930) prepared a form following a technique developed by Ream (1924). They used simple language and referred specifically to many school activities. Their questionnaire has eight sections, as follows:

	No. of Items
1. Occupations	68
2. Activities	24
3. School Subjects	20
4. Job Activities	20
5. School Activities	41
6. Prominent Men	12
7. Things to Own	26
8. Magazines	23

All items are to be checked for Like, Indifference, or Dislike. Three keys are furnished which yield separate scores for interests similar to those for students in academic, commercial, and technical courses. The odd-even reliability for each scale is above .86. The scales were constructed by comparing preferences of students in a particular curriculum with students in general. From these comparisons stencils were made which allotted points for the answers showing large differences between groups.

3. WALLAR AND PRESSEY'S OCCUPATIONAL ORIENTATION INQUIRY. Another interesting self-rating scale is that of Wallar and Pressey (1937). This consists of 222 names of occupations which are to be rated from four important standpoints: knowledge from experience, enjoyment, ability, and opportunity. Students are asked to use a four-point scale and to rate all items for the first aspect, knowledge, before rating the second aspect, enjoyment (see Table 79). This blank showed that among 444 college freshmen the average student reported 5 occupations about which he had some knowledge from firsthand experience. There were 24 occupations which he was sure he would enjoy and 14 for which he believed he possessed above-average ability. Seventy-eight per cent of the group stated that they had no acquaintance or contacts through which they believed they might secure work of any kind.

This form also requires a written vocational history and a self-evaluation of one's present situation, so that the items checked may be compared with a record of development.

No data have been reported which show the intercorrelations among ratings of knowledge, enjoyment, and ability. Although these correlations will probably be positive, one might expect them to be rather low, since the factors involved appear to be somewhat independent of one another, and since opportunities to get firsthand knowledge of types of work are usually very limited.

4. STRONG'S VOCATIONAL INTEREST BLANK. Strong published his first edition in 1927 and a revised edition in 1938. He followed very closely the methods of Ream (1924) and Freyd (1922). A large variety of names of occupations, amusements, school subjects, types of people, activities, and self-ratings of habits or traits were selected for their power to indicate differences in interests among a large variety of

TABLE 79

Occupational Orientation Inquiry

All of us know that many times vocational choices are based on inadequate reasons. Perhaps there is a passing interest in that field or there may be a very good friend in that occupation. However, in choosing an occupation one should carefully analyze all aspects of the work in any job which he might care to choose as his life career. Hence, you are now asked to consider the occupations listed below from four important standpoints. Look the list over carefully and, by means of the scales below, rate yourself in each occupation as to your knowledge, interest, ability, and your probable opportunity for placement in each field.

Knowledge

4. I have been regularly employed in the job.
3. I worked part time or helped in the job.
2. I have general knowledge of job but no knowledge of details.
1. I have less than average knowledge of the job.
   If you have no acquaintance with the job, leave the space blank.

Interest

4. I think I would enjoy it as a life work.
3. I think I could develop sufficient interest to enjoy it.
2. I am indifferent but would work at it if there were the chance.
1. I would work at it only as a last resort.
   If you would not work at it under any conditions, leave the space blank.

Ability

4. I think I have exceptional ability in the field.
3. I could be successful in the job.
2. I would be able to do average work.
1. I would find it difficult to work in this field.
   If you would find it practically impossible to make a living at it, leave the space blank.

Opportunity for Replacement

4. I am practically assured of a job.
3. I have contact with the field.
2. I know people in the field through whom I might gain contact with the job.
1. I have no existing contacts in the job.
   If you have almost no possibility of contact in the job, leave the space blank.

In other words, as regards each occupation, indicate your knowledge, interest, ability, and opportunity in this way: 4 means much; 3, above average; 2, average; 1, below average; and a blank space means very little. Work rapidly as your first impressions are valuable. Begin at once.

Knowledge	Interest	Ability	Opportunity		Knowledge	Interest	Ability	Opportunity		Knowledge	Interest	Ability	Opportunity	
				1. Accountant					31. Cashier					61.
				2. Actor					32. Caterer					62.
				3. Advertiser					33. Cement mixer					63.
				4. Animal husbandryman					34. Chauffeur					64.
				5. Army officer					35. Chemist					65.
				6. Artist					36. Civil service employee					66.
				7. Artist's model					37. Clergyman					67.

(Wallar and Pressey, 1937. By permission of the Editor. *Occupation.*)

occupational groups. It was found that the average high-school and college student would check the total blank which contained 400 items in about forty minutes, indicating Like, Indifference, or Dislike on each item, or choosing between two or more items.

This test was standardized from the scores of groups of men who were considered to be successful representatives in a particular vocation. These men, whose average age was about 43 years, had been engaged in one occupation for at least three years. Most of these groups included from 500 to 1,000 men whose average income was $2,500 a year or more. The scores are thought to be significant for ages from 20 to 50 years. Before 20 years of age, changes of interest are sometimes large.

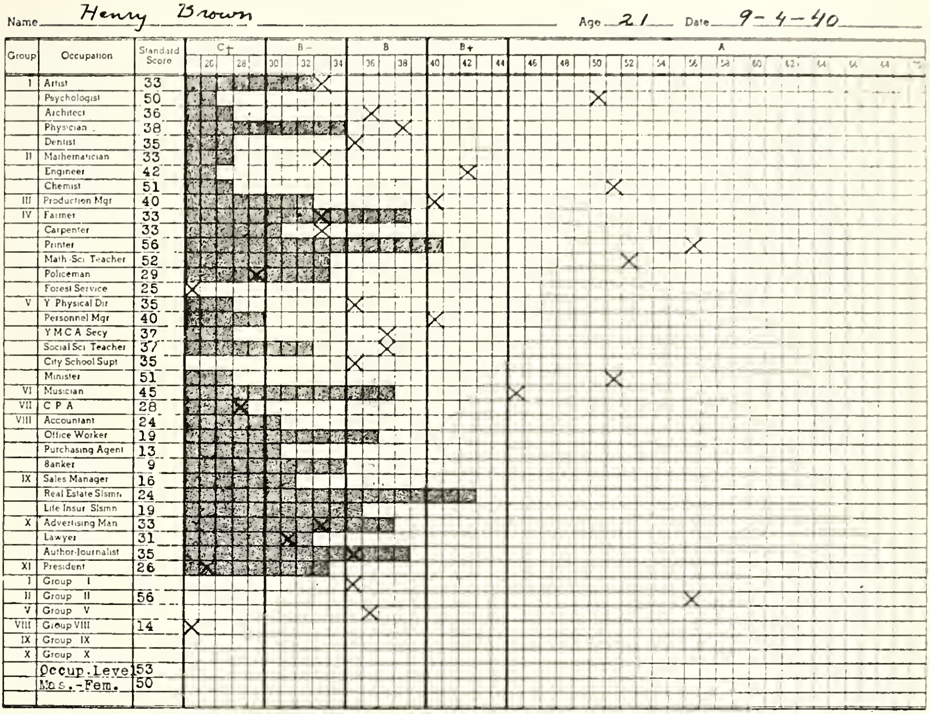
Scoring keys for Strong's Vocational Interest Test for Men are now available for the 32 occupations listed in Table 80 and for 6 general groups of occupations. The procedure for making a scale for each group was first to calculate the percentage of the group which marked each item Like (L), Indifference (I), or Dislike (D). Similar percentages were found for a large men-in-general group. The differences between the special and the general groups were found for each item. For instance, in calculating the weights to be assigned in the first item, "actor," when the interest of personnel managers was to be evaluated, the per cents of personnel managers who checked each response were compared with the per cents of the general group, thus:

	L	*I*	*D*
Personnel Managers	49	38	13
All others	38	35	27
Difference	+11	+3	−13

The differences were transposed into smaller figures for ease of handling so that the final weights for this item were: L = +2, I = +1, and D = −3. An individual's score is the sum of his plus and minus marks on all items. If his score

TABLE 80

Report on Vocational Interest Test for Men



Name *Henry Brown* Age *21* Date *9-4-40*

Group	Occupation	Standard Score
I	Artist	33
	Psychologist	50
	Architect	36
	Physician	38
	Dentist	35
II	Mathematician	33
	Engineer	42
	Chemist	51
III	Production Mgr	40
IV	Farmer	33
	Carpenter	33
	Printer	56
	Math-Sci Teacher	52
	Policeman	29
	Forest Service	25
V	Y Physical Dir	35
	Personnel Mgr	40
	Y M C A Secy	37
	Social Sci Teacher	37
	City School Supt	35
	Minister	51
VI	Musician	45
VII	C P A	28
VIII	Accountant	24
	Office Worker	19
	Purchasing Agent	13
	Banker	9
IX	Sales Manager	16
	Real Estate Slsmn	24
	Life Insur Slsmn	19
X	Advertising Man	33
	Lawyer	31
	Author-Journalist	35
XI	President	26
I	Group I	
II	Group II	56
V	Group V	
VIII	Group VIII	14
IX	Group IX	
X	Group X	
	Occup. Level	53
	Mas.-Fem.	50

(Strong, 1938. By permission of Stanford University Press.)

falls within the highest 69 per cent of the scores of personnel managers, he is given an "A" rating in this field. If his score falls within the next lowest 29 per cent of the scores, he is given a "B" rating. If it is lower than the score obtained by the lowest 2 per cent of the personnel-manager group, he is given a "C" rating.

The "A" rating indicates marked similarity of interests with those of persons successfully engaged in an occupation. The "B" rating shows some similarity and the "C" rating, no similarity or contrary tendencies. An individual's score may also be transposed into standard scores or centiles from norms, such as those for engineers in Table 81. The blank must be scored for each occupational scale. Hence, if scores are desired for 32 occupations, it is necessary to score the blank 32 times. The labor involved is large whether the scoring is done by hand or by machinery. A report blank, Table 80, gives a profile which indicates both strength and direction of interest. Three uses for these profiles are suggested by Strong: in the hiring of employees, in the admission of students to college, and, chiefly, in aiding a person to decide upon which occupation to enter. A low score in interest of musicians, for instance, is taken to mean that one would not be satisfied in this occupation. A high score presumably means that one's satisfaction will be as great as most of those in the profession. If one has several high scores, some occupation may be found which will combine these interests.

Strong (1938) reports fairly high reliabilities. The mean correlation of odd-even items was .87 for the thirty-six revised scales based on records of 285 Stanford seniors. The retest correlation with an interval of one week averaged approximately .869. The mean correlation on twenty-one occupational scales over a five-year period was .75 for a group of 285 college men. This is an exceedingly high figure for retests of this sort.

Strong believes the test has great validity because of three sorts of evidence. First, it distinguishes quite clearly those who are successfully engaged in one occupation from men in general. Table 81 shows a very large difference between scores of successful engineers and scores of Stanford students. Only 15 per cent of 933 nonengineering men rated an "A" in engineering interest, and of this 15 per cent many were in occupations which applied physical and mathematical concepts. Second, validity is indicated by the degree to which interest scores corresponded to success in occupations. Among 181 life insurance agents it was found that 67 per cent of those with "A" ratings wrote at least $150,000 of policies a year, and less than 6 per cent of men with "C" ratings achieved this result. A third evidence of validity comes from records showing that men who continue in an oc-

TABLE 81

Engineering Interest Scale Norms

513 Engineers			Percentile Scores		
Raw Score	*Standard Score*	*Rating*	*Engineers*	*306 Stanford Freshmen*	*285 Stanford Freshmen*
220	76	A	99		
210	74	A	99		
200	71	A	99		
190	69	A	98		
180	67	A	96	99	99
170	64	A	93	99	99
160	62	A	89	99	98
150	60	A	83	98	96
140	57	A	75	97	95
130	55	A	66	95	94
120	53	A	57	92	92
110	50	A	48	88	88
100	48	A	42	87	83
90	45	A	33	83	80
80	43	B +	24	79	76
70	41	B +	17	75	73
60	38	B	14	69	68
50	36	B	9	63	66
40	34	B −	6	59	62
30	31	B −	4	53	54
20	30	B −	2	48	44
10	27	C +	1	42	41
0	24	C +	1	37	35
− 10	22	C	1	31	27
− 20	20	C	1	25	22
− 30	17	C		20	18
− 40	15	C		14	12
− 50	12	C		10	9
− 60	10	C		7	7
− 70	8	C		5	5
− 80	5	C		2	2
− 90	3	C		1	2
− 100	1	C		1	2
− 110	− 2	C			1
− 120	− 4	C			1

(Strong, 1938, p. 10. By permission of the Stanford University Press.)

cupation obtained higher interest scores in it than in other occupations, and that men changing from occupation A to occupation B, tended prior to the change to score higher in B than in A.

Estes and Horn (1938) found remarkable differences in engineering interest scores among groups of engineering students in five curricula. The students who were studying the mechanical engineering curriculum had a median score of 208, which was approximately the 99th centile in Strong's engineering group. On the other hand, the students in civil engineering averaged only 39, and in chemical engineering, 18. These scores would be similar to those of the lowest 3 or 4 per cent of Strong's engineers.

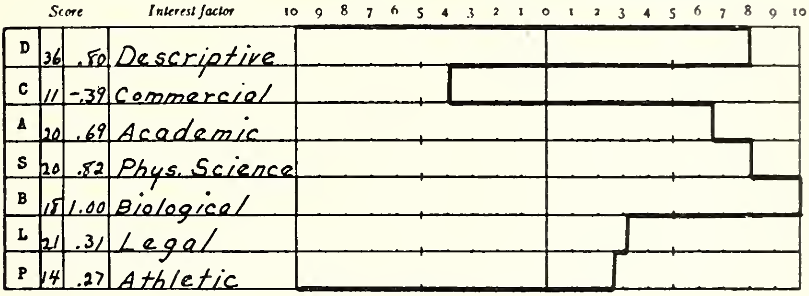
In order to evaluate the interests of these five groups, five new scales were constructed following Strong's technique. Then it was found that when students in one curriculum were scored on the scale of another curriculum 61 per cent were rated either "B−" or "C." The authors also calculated the correlations between the Estes-Horn scales and the Strong engineering scale. The scales for electrical and mechanical curricula groups correlated approximately .70 with Strong's engineering scale. The correlation with the chemical engineering interest scale was −.27; with the industrial, −.63; and with the civil, −.09. Since chemical and industrial engineers had not been included in Strong's criterion group, there is some justification for the negative correlations. Strong did, however, include civil engineers in his group; hence the zero correlation is hard to explain. These results probably indicate that interest patterns vary so much among engineering students that a single scale will not suffice to measure their interests. Only the electrical and mechanical engineering students showed patterns of interest which were similar to Strong's sample of engineers of approximately 43 years of age.

5. Comparison of Inventories. A thorough comparison of inventories to show their relative merits is not feasible at this time, but for purposes of illustration the records of a college student, Henry Brown, on four self-rating interest inventories are available. Table 80 shows his scores on the Strong test. His highest interests were like those of a Printer, Psychologist, Chemist, Mathematics or Science

Teacher, and Minister. Moderate interest scores are found for Engineer, Production Manager, Personnel Manager, and Musician. He shows a major interest in Group II, Physical Sciences, and moderate interest in Group I, which has not been named, and Group V, Personnel and Social Occupations. Table 82 shows his scores on the Thurstone Vocational Interest Schedule, which consists of 89 names of occupations to be checked. He shows unusually high interests in Biological and Physical Sciences and in occupations which deal with personal social situations. His interest in Commerce is low, and in Legal, Athletic, and Academic Fields, moderate. Table 83 shows his profile on Kuder's (1939) Preference Record. Kuder employed 330 two-choice items concerning preferences for many sorts of recreation, work, people, and clothes. Henry Brown shows very high scores for Scientific work, moderate for Social Service, average for Computational, Artistic, Literary and Persuasive occupations, and a small score for Musical activities. Table 84 shows the ten scores of Cleeton's (1937) Vocational Interest Inventory. Henry Brown's interests were found to be very high in Biological Science, Sales, Physical Science, Social and Counselling work, Legal and Literary activities, Mechanical Trades, and Artistic and Musical skills. His interest scores

TABLE 82

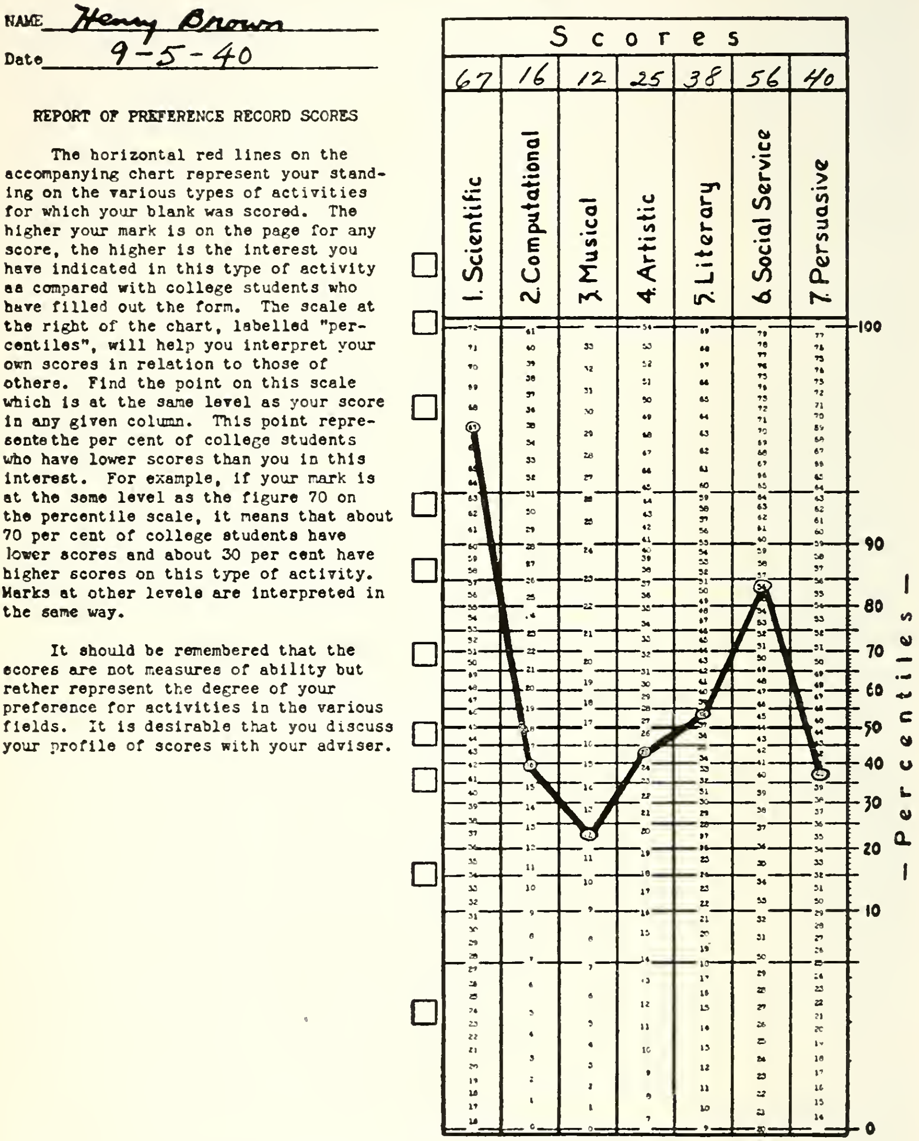
Vocational Interest Profile of Henry Brown



(Thurstone, 1935. By permission of the Psychological Corporation.)

TABLE 83

Preference Record



NAME Henry Brown

Date 9-5-40

REPORT OF PREFERENCE RECORD SCORES

The horizontal red lines on the accompanying chart represent your standing on the various types of activities for which your blank was scored. The higher your mark is on the page for any score, the higher is the interest you have indicated in this type of activity as compared with college students who have filled out the form. The scale at the right of the chart, labelled "percentiles", will help you interpret your own scores in relation to those of others. Find the point on this scale which is at the same level as your score in any given column. This point represents the per cent of college students who have lower scores than you in this interest. For example, if your mark is at the same level as the figure 70 on the percentile scale, it means that about 70 per cent of college students have lower scores and about 30 per cent have higher scores on this type of activity. Marks at other levels are interpreted in the same way.

It should be remembered that the scores are not measures of ability but rather represent the degree of your preference for activities in the various fields. It is desirable that you discuss your profile of scores with your adviser.

Scores						
67	16	12	25	38	56	40
1. Scientific	2. Computational	3. Musical	4. Artistic	5. Literary	6. Social Service	7. Persuasive

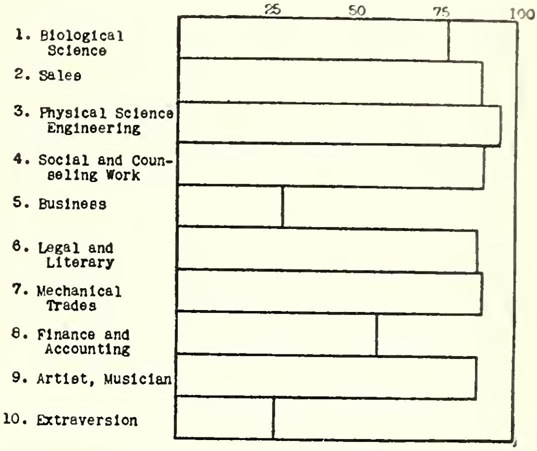
(Kuder, 1939. By permission of the author.)

were low for Business and Financial work, and he showed a low Extroversion Index. Cleeton's inventory is composed of items much like those of Strong, but the scoring was based on job analyses as well as group preferences.

These four reports agree in indicating a very high preference for work in Physical Sciences. The Thurstone, the Kuder, and the Cleeton scales show high ratings in Biological Sciences; the Strong blank has no corresponding category. Kuder's scale contains no Sales or Commercial ratings. The other three scales rate him low for Business or Commercial interests. On Sales interests he received a high rating on the Cleeton, and a low rating on the Strong. The other two authors have no Sales category. High ratings for occupations which deal with Personal Adjustment were found on the Thurstone and Cleeton inventories and moderate ratings on the other two. Artistic interests were marked high in all except the Thurstone scale, where no Artistic category is found. Both the Kuder and the Strong inventories show low scores in Musical interests.

TABLE 84

Cleeton's Vocational Interest Inventory of Student, Henry Brown



(Drawn from data furnished by Gertrude Muxen, University of Michigan, Bureau of Appointments and Occupational Information.)

One must conclude that the four inventories correspond fairly well on similar categories, and that the inconsistencies may probably be explained by the somewhat different selection and grouping of items. All four inventories seem to require about the same length of time for administration, approximately 40 minutes. It is not easy to compare scoring costs because the inventories yield different results. The Strong blank yields more separate scores than the others, but some of these correlate fairly highly with each other. For most purposes a few broad fields of interest are sufficient.

An interesting comparison of technical information and interests is provided by further reference to Figure 33, page 210. Henry Brown's vocabulary scores gave him the following centile ratings in a large third-year college group:

Physical Sciences	98	Commerce	67
Mathematics	95	Government	67
Biological Sciences	87	Fine Arts	43
Human Relations	84	Sports	38
	Total Score	91	

These figures show that his three highest information scores were in his three highest interest fields, and that his four lowest information scores were in his medium or low interest fields. His total vocabulary and his total interests were both much higher than average. The correspondence between interest and information tests has been shown to be fairly high by McHale (1930) and Flanagan (1939) when opportunities for acquiring information were nearly the same for all persons in the group. However, more research is needed to show what factors are involved in learning and retaining technical information and in developing interests.

6. FACTOR ANALYSIS OF INTERESTS. Thurstone (1931) applied a multiple-factor analysis to the results of Strong's Vocational Interest Tests for Men. From the eighteen scores secured for each of 237 persons, four main factors appeared. These seem to correspond fairly well to interests in science, language, business, and people. He pointed out that the variance of the scores was not fully accounted for by these four factors, and conducted another study based on the Thurstone Vocational Interest Schedule. This consists of 89 names of occupations which are to be checked to show a person's like, indifference, or dislike toward the occupation named. A factorial analysis of these 89 items yielded seven independent factors, which were tentatively identified as:

1. *Descriptive* (of persons and social situations)
2. *Commercial*
3. *Physical Science*

4. *Biological Science*
5. *Legal*
6. *Athletic*
7. *Academic* or *literary*

He also suggested an eighth factor of interest in art, which was not clearly indicated in his study because only a few items referred to artistic activities. Table 82 shows an individual profile based on these factors from the Thurstone Vocational Interest Blank.

In 1934 Strong made a factorial analysis of his own blank using both men and women. The results yield five general factors tentatively called an interest in people, business, intellectual activities, science, and language. Interest in people also corresponded to feminine interests to a marked degree. On the basis of similar factor patterns, Strong grouped the occupations into eleven categories, as shown in Table 80.

The Strong blank was applied to 133 boys in high school by Carter, Pyles, and Bretnall (1935). The boys ranged from 12 to 19 years of age with a median of approximately 16. Using Strong's scales, twenty-three scores were obtained for each boy. These were correlated with one another and with age. The correlations with age ranged from −.11 for journalist to .33 for purchasing agent, median .06. The intercorrelations of interest scores proved to be similar to those found for college men, with the exception of four scales—those for Minister, Y.M.C.A. Secretary, School Man, and Personnel Manager. Previously, Strong found these four scales to be more highly correlated with maturity of interests than the other scales. Factorial analysis of the correlation matrix by Thurstone's methods yielded three factors which seem to correspond fairly well to those found by the other investigators — namely, interests in persons, in science, and in language. A fourth factor, not so clearly isolated, seems to correspond to interest in business. The authors point out that the sum of the squares of the loadings had a mean value of .83 for seventeen scales. These same scales had previously reported a mean reliability of .88. The small difference between these figures indicates that specific factors are very small in the different scales.

In spite of the similarity of factors shown in the last two studies, the factor loadings for the various professional scales are not very similar. The specificity of items shown by factorial analyses is greater for groups of men than for boys. These results doubtless are due to the different experiences of the groups tested. The interpretations of items by boys are probably more vague and also more romantic and adventuresome than the interpretations by men.

Dwyer (1938) found that nineteen of the Strong occupational scores could be calculated from only four scores. If the scores of a person were known in the field of a physicist, a minister, a life insurance salesman, a journalist, then all the other scores with the exception of those for certified public accountant and farmer could be calculated with remarkable accuracy.

Recently Crissey and Daniel (1939) have reported a factorial analysis of eighteen scores on Strong's Vocational Interest Blank for Women. By using Thurstone's centroid procedure, four factors were found which accounted quite well for their correlation matrix. They tentatively named these factors:

1. *Male Association:* housewife, nurse, secretary or stenographer, and general office worker
2. *Interest in People:* Y.W.C.A. secretary, lawyer, and teacher of social sciences
3. *Interest in Language:* teacher of English, teacher in general, librarian, and author
4. *Interest in Science:* physician, dentist, teacher of mathematics, and teacher of physical science

One must conclude from these factorial studies that the most valuable analyses will come from work which appraises each item in a battery separately. When a number of items are added together, as in the scales described above, the total score will usually not represent a single variable but a number of variables in unknown proportions. Factorial analysis of these data, then, is dependent to a large degree upon the particular selection of items and scores made by the author, and only indirectly upon the true relationship of the items as shown by their coexistence in the same person. Factorial analysis of Strong's preference scores will usually give different results in different groups, since the items themselves have variable meanings. Part of the difficulty is due to the failure of questionnaires to give adequate descriptions of vocations. A check list of vocational names such as banker or engineer will be variably interpreted by persons according to their own experiences. A very worth-while approach to this problem is the use of questionnaires which ask a person to indicate preferences for types of activities rather than the names of occupations. A thorough study of this sort has not yet come to hand, although many forms have been prepared which include some activity preferences, particularly Stewart and Brainard's (1932) Specific Interest Inventory.

III. PRACTICAL RESULTS

The application of scales such as those just described has yielded a number of findings of considerable interest. These will be discussed under the headings, permanence of interest, prediction of scholastic success, and prediction of vocational success.

A. Permanence of Interest

Pioneer work by Thorndike (1912) and (1917) showed that there are fairly high correlations among rank order of interests in elementary school, high school, and college when studied in retrospect. Each college student was asked to show his preference during each period for the following seven subjects: mathematics, history, literature, science, music, drawing, and handiwork other than drawing. The median correlation between rank assigned to elementary school and high school periods was .85 among 444 college students. The correlation dropped to .79 between high school and college periods, and to .66 between elementary school and college.

Thorndike (1935) made a retrospective study of changes of interest with age among 141 adults in their fifties. About equal numbers of clergymen, men teachers, women teachers, lawyers, psychologists, and noncollege trained workmen and storekeepers were asked to rate amount of liking or disliking for 17 types of activities using a ten-point scale. Typical likes for each decade were to be rated separately. The list of activities was:

1. Reading fiction
2. Reading non-fiction (except newspapers)
3. Reading newspapers
4. Outdoor sports (not competitive)
5. Outdoor competitive games
6. Sedentary games
7. Dancing
8. Playing a musical instrument
9. Listening to music
10. Theater or movies
11. My regular job
12. Politics
13. Welfare work
14. Talking with old friends
15. Making new acquaintances
16. Traveling and seeing new places
17. Sheer idleness

Thorndike found that median changes between the decades 20 to 29 and 50 to 59 were small for all of the groups. There were slight decreases in liking for outdoor sports and competitive games, dancing, and reading of fiction, and slight increases in reading nonfiction and newspapers. The clergymen alone reported small increases in liking for their regular jobs, welfare, music, old and new friends. Thorndike believed that these ratings were fairly free from large errors due to memory, or errors due to systematic variations in the use of the rating scale. He concludes that average likes do not show marked changes over these 30 years except for interest in strenuous physical activity.

A study by Buck (1936) used the Pressey Cross-Out Test for evaluating likes, worries, and wrong acts among 4,000 students from the seventh grade to college. Each of the three parts of the test consists of 25 rows of five words each. Each row contains one word which is supposed to cause a reaction of indifference, and four words which usually are emotionally toned. Thus, a row of words to be marked as *wrong to do* is

Begging, smoking, flirting, spitting, giggling

Students were requested to cross out all wrong acts and to encircle the worst in each row.

The results show marked reductions in early adolescent interests, such as circuses and ferris wheels, and increases in adult interests, such as business and cards. Childhood worries about fire and lightning were replaced by the adult worries about family, religion, and marriage. There was also marked decrease in the things considered wrong. Smoking, flirting, borrowing, and divorce were considered wrong by most children, but by few adults. The report also compares college samples in 1923 and 1933. The ten-year period showed a marked drop in things considered wrong, but nearly the same results for likes and worries.

Strong (1931) published a report of changes of interest with age based on one application of the Strong blank to 2,340 men between the ages of 20 and 60 in eight professions. In every case the results in all eight occupational groups were similar. He found that the total number of likes was nearly the same, 34 per cent for each decade. The most noticeable changes occurred in preferences for persons and amusements. "Driving an auto," "musical comedy," "explorer," "orator," "yell leader" all declined as preferable activities, while "nights at home," "art galleries," "contributing to charity" all increased. In general, the activities liked most by 25-year-old men were liked a little better by older men, and the things liked least by 25-year-olds were liked less and less with older groups. The reverse was true of preferences for occupations and school subjects. Slight decreases were noted in the per cents of men liking occupations, including one's own occupation.

A number of follow-up studies compare two self-ratings of interest at various intervals. Columba (1926) found that only 46 per cent of 244 children in Grades VI to VIII named the same school subjects as the one they liked best on two questionnaires a year apart. Willett (1918) found that only 26 per cent of boys and 36 per cent of girls in a high school preferred the same school subject on two occasions a year apart. These same children showed only 8 per cent permanence of interest for boys and 18 per cent for girls over a two-year period.

An able review of studies of permanence of vocational interests by Fryer (1931) shows varied results due to differences in technique. In general, greater permanence is found when interests are reported for a few broad categories than for many specific occupations. Furthermore, self-estimates of interest in an occupation do not correspond very highly with scores from inventories, but they seem to have about the same permanence. Studies of school children usually employ estimates, whereas studies of adult preference have commonly used inventories.

Fryer found that nine studies of first choices for vocations among elementary-school children showed a mean of 46 per cent identity between estimates one year apart. A three-months' course of instruction dealing with occupations previous to the first questionnaire tended to raise the permanence of choices considerably. A study of prime interest in high-school groups showed a mean of 57 per cent identity for an interval of one year, 45 per cent for two years, 25 per cent for three years, and 26 per cent for four years. Among college students a one-year interval yielded 83 per cent of similar choices; and a two-year interval, 56 per cent. The per cents quoted are not strictly comparable, but they show a tendency for greater stability of choice among the more advanced groups.

Permanence of inventory scores for vocational interests have been reported using several fairly elaborate forms. Total score for both mechanical and social interests taken fifteen months apart were found by Hubbard (1928) to correlate approximately .40 among 150 high school students and .51 among a similar group of college students.

Strong (1934) (1935) reported higher correlations for college groups than for high school groups. A retest of Stanford University seniors after five years resulted in a mean correlation of .75 between scores on the same scales. This finding was more remarkable in view of the fact that one-half of the group had changed their occupation within the five years. The change, however, was generally in line with high interest scores which had been recorded in college. The scores at age 27 were slightly higher than at age 22. Strong believes that this is due to the fact that the scales were constructed from scores of men whose average age was 43 years. Strong also reports that 61 ministers showed a retest correlation of .75 and 32 certified public accountants, one of .81 upon two tests separated by eighteen months.

B. Prediction of Scholastic Success

Thorndike (1912) (1917) (1921) reported correlations between students' self-ratings of interests in seven educational subjects and their self-ratings of abilities in these same subjects. He found correlations by a rank-order method of approximately .89, and concluded that interests were highly predictive of abilities. King and Adelstein (1917) reported that on two different occasions they found correlations of approximately .73 between self-ratings for interest and for ability. Fryer (1927) made a similar study of college students and reported correlations of approximately .60 when the number of academic subjects was not limited. These correlations are probably spuriously high because they are all self-ratings. Self-ratings usually cannot be made with a marked degree of isolation in situations such as these. For instance, if one rated himself as enjoying a school subject very much, he might, on a later occasion, rate himself as very able in this subject on the basis of a few successful episodes. If he rated himself as dissatisfied, he might also remember a few failures and therefore rate himself as being inept. The same sort of logical error may occur if ranking for ability precedes ranking for enjoyment.

Terman (1925) compared self-ratings of interest in school subjects with teachers' estimates of ability in the same subjects, among 527 normal and 643 gifted children with IQ's of 140 or higher. The separate correlations for boys and girls in each group averaged approximately .417.

Definite checks of self-ratings of interest with school achievement are reported by several authors. A study by Columba (1926) compared the self-rankings of interest in six school subjects with scores on the Stanford Achievement Test. The subjects were arithmetic, reading, spelling, geography, history, and language. The study was conducted among 356 pupils in the fifth, sixth, seventh, and eighth grades. In no case did the persons who claimed that they especially preferred a subject have a significantly higher mean achievement than those who did not make such a claim.

Kornhauser (1929) applied a general interest inventory to 110 college freshmen, and then made a scoring key which distinguished the highest and lowest quarters of students on the basis of their grades in all subjects. Correlation between general scholastic interest as determined from this scoring key and grades was found to be only .17 on a small sample of students.

A similar study made by Shuttleworth (1929) classified

269 women and 347 men into tenths on the basis of freshman grades and a general intelligence test. Then, from a 200-item general interest blank, 34 items were found which differentiated the highest and the lowest tenths of the group when divided on the basis of intelligence. Sixty items were found which distinguished between the highest and the lowest scholarship tenths. A new form was devised which was found to correlate .43 with the first semester grades and .61 with a comprehensive college entrance examination. Shuttleworth further developed separate scores for interest in six kinds of activities:

1. Intellectual and cultural
2. Spiritual, ethical, and religious
3. Freedom from conventions
4. Political liberalism
5. Freedom from pleasure seeking
6. Nonmechanical

He found rather low correlations between scores in these fields, median, .23, which may indicate either independence or a low retest reliability due to the small number of items and persons which he employed. The scores for intellectual interests and freedom from conventions showed a slightly higher relationship with grades and mental tests than the scores in other fields of interest.

Another careful study of the relationship between interest and college grades was made by Jacobsen (1928) on the basis of a preliminary application of a revised form of the Carnegie Interest Inventory (1921 edition). He selected 66 items which differentiated 133 women who were superior in both scholarship and intelligence tests from 94 women who were inferior in scholarship but superior in intelligence. When applied to a control group of 304 freshman women, this scale correlated .19 with intelligence test scores and .32 with grades in the freshman year. A subsequent refinement of this scale correlated .21 among men and .40 among women with school grades, and .30 among men and .44 among women with intelligence tests. The combined prediction of grades from interests and intelligence was .58 for men and .63 for women. These figures indicate that when interest scores are added to intelligence test scores, the prediction for freshman grades is raised by about .10.

Studies of the sort just described fail to compare special interests with special achievement, and hence they have not proved to be very discriminating. They do indicate that differences between scholastically superior and inferior students exist, but that differences in interest between various occupational groups are large and important. Several studies have been made of interests in particular fields.

Cowdery (1926), using a small group, developed scoring keys which distinguished among students in engineering, law, and medical work. The correlations between grades in professional school and interest scores were for medical students, .01; for engineering students, .34; and for law students, −.12. Another study by Remmers (1929) reported zero correlations between grades in agriculture or engineering and interests on the Purdue Interest Report Blank.

Garretson (1930) found only insignificant correlations between academic interests measured by his blank and academic grades in high school and scores on Terman's Group Test of Mental Ability. Between commercial interest and grades in commercial subjects the correlations were also zero. Between technical interest and grades in technical subjects the correlation was .29.

The last seven of the reports just discussed generally failed to show significant relationships between self-ratings of preferences and scholastic grades. This does not mean that significant relationships were lacking. Fryer (1927) pointed out that although self-ratings result in a ranking of academic subjects for each individual, the assignment of school grades does not show the individual's relative success on various subjects but only his position in a larger group. When Fryer eliminated this complication by using individual ranks of grades, he found rank-order correlations with preferences ranging from .60 to .80. A similar study by Greene (1939) reported rank-order correlations ranging from .65 to .79 among 463 college sophomores, between preferences and technical vocabularies.

C. Vocational Predictions from Interest Scores

A widely accepted belief is that enjoyment of a type of activity plays a dominating role both in the selection of an occupation and in continuous and successful employment. This belief is neither widely confirmed nor denied by the present results. Many persons doubtless succeed in a vocation in spite of early dislikes. Reasons for occupational choices often include financial and religious considerations. One would not expect, therefore, that a mere statement of probable enjoyment for a type of work, which a person may never have tried, would clearly predict success in after years. In order to evaluate the part played by interest in vocational success, one would have to follow a large group of persons over a period of years, and to note in what ways and by what means likes are changed. Data of this sort are very meager. The ten-year follow-up study reported by Thorndike in 1934 did not include appraisals of interests in 1922. Most of Strong's work compares interest scores with success at the time of the test, but one study published in 1934 reports a five-year follow-up of 223 college seniors. This group was selected in 1927 from approximately 400 men on the basis of their willingness to fill out the blanks, and five years later by availability as well as willingness to co-operate. For each person the results included two sets of interest scores, one in 1927 and one in 1932; a definite choice of occupation in 1927 (for 79 per cent of the group); and records of actual employment in 1932. These scores were converted into ranks for each student.

The results showed that 48 per cent of the group had continued for five years in the occupations which they had originally chosen, 9 per cent had made a slight change in occupation, 22 per cent had made a marked change, and 21 per cent had changed from *undecided* to a specific occupation. A slight change was judged to have occurred when one shifted from psychiatrist to psychologist or from law to journalism. A marked change was illustrated by a turning from law to farming or from chemistry to real estate sales.

From the 1927 test it was found that among those who had continued in their chosen field the interest score in the chosen field had an average rank of 1.78. This was the highest average rank of any interest score, and it indicated that continuation in an occupation can be predicted with considerable accuracy from interest scores in this group of students. For those who had changed their occupations, the mean rank of the occupation to which they changed was 6.0. This means that on the 1927 test there were other occupations which were ranked ahead of that which was finally selected. Among those who had no definite choice in 1927, the rank of the occupation which they had entered by 1932 was 4.38. These figures show that even when there was a lack of decision or a later change, the interest scores in 1927 were somewhat predictive of the field into which a person would go. These predictions would doubtless be higher if the interest blank had yielded scores for fairly broad classes of occupations, such as the scores which are now available from Strong's material. An interesting addition to this study would be the inclusion of measures of job satisfaction of an analytical sort. It may be that persons enjoy some parts of their work much more than others and that the enjoyable aspects would correlate highly with interest scores. The prediction of vocational choice or vocational success from interest scores alone is likely to be misleading unless other factors, such as ability and training, are taken into consideration, since these factors undoubtedly contribute both to success and to enjoyment.

IV. FUTURE RESEARCH

The tests which have been described and the reports which summarize work in this field seem to show very definitely that valuable additions to prediction in academic and vocational success can be secured by appraisals of motives. These studies also show a lively interest in the field and the need for a number of refinements. Refinements can be made along the following lines:

1. Item analysis: Nearly all workers in this field have noted that the interpretation of items is often beset with ambiguities. Scales which use more descriptive items and more discriminating rating methods will undoubtedly reduce ambiguity to a marked extent.
2. Factorial analysis: When unambiguous items are used it is possible to apply statistical techniques to show the relationships of these items in the experiences of various people. Factorial analyses will be of value in establishing the independence of items, and hence in constructing scales which will yield significant profiles. Factorial analyses will also shed some light on the nature of the organization of wants and drives in various social and age groups.
3. Sampling the population: Much research is needed to show the extent to which means scores of a group are typical of the most successful persons in the group. A more analytical set of measuring devices will show what aspects of work are most enjoyable among those who have achieved various degrees of success. Much research is also needed to show essential differences in interests of various age groups.
4. Reduction of misrepresentation: Appraisals of motives are sometimes subject to intentional or unintentional misrepresentation. Part of this may be due to a desire to seem socially or vocationally acceptable. Several in-

direct techniques for the measurement of motives offer possibilities of reducing misrepresentation.

5. Follow-up studies: Although some general tendencies for persons to change with age and experience are fairly clear from retrospective studies, yet much needed information will come from follow-up studies. These would give facts concerning the permanence of various types of interests, the types of persons who fluctuate the most, and the effects of counseling and work experience on satisfactory adjustments.

16

APPRAISALS OF ATTITUDES

I. INTRODUCTION

There does not appear to be a clear-cut distinction between likes and dislikes, discussed in the previous chapter, and attitudes, discussed here, but a rough practical discrimination can be made on two bases, subject matter and ethics. The subject matter of a preference inventory is largely limited to personal activities, whereas attitudes involve broad questions of policy or values for a group. In ascertaining personal preferences, one is usually asked to state his real satisfactions regardless of their moral significance. A person may like some activity which he believes is immoral. In attitude scales, however, a person is required to voice his approval of institutions, activities, races, or principles. His personal activities are not emphasized as much as his opinion concerning what should be done for the good of all.

A simple type of unscaled appraisal of attitude is a single vote or judgment for or against some act. Voting activities range from selecting a beauty queen to approving a far-reaching governmental reform. Voting is often considered to be a superficial method of evaluating attitudes, since many persons cast votes without understanding a proposal. Often those who are least informed vote with most assurance; then, a few minutes later, after some dramatic appeal, they are ready to change their votes. However, when persons are adequately informed and uninhibited, voting is an excellent means of attitude appraisal. Practically, voting is designed to decide some present issue, and rating scales are designed to evaluate usual goals or preferences. However, there seem to be no important psychological differences between voting and the rating scales discussed below. The scales represent attempts to refine voting.

II. TYPICAL SCALES

Among the outstanding studies in this field are Watson's (1925) *Test of Public Opinion,* Thurston and Chave's (1929) *Scales of Attitudes,* Allport and Vernon's (1931) *Study of Values,* and Murphy and Likert's (1938) *Study of Public Opinion.* Two scales which illustrate interesting techniques but which have as yet been little used are Horowitz's (1936) measurement of attitude toward the negro, and Brandt's (1937) analysis of replies to questions of personal ethics. These scales will first be described, and then some typical results will be quoted.

A. Watson's Test of Public Opinion

Watson's Test of Public Opinion consists of six parts of approximately 50 items each. Each part uses a different method of eliciting opinions. The first part is a *cross-out* test of 51 items in which one is instructed to cross out every word which is more disagreeable or annoying than agreeable or attractive. The second part is called a *degree of truth* test. Here a person is asked to rate 53 items on a five-step scale, from utterly true to utterly false. The first five items are shown in Table 85. The third part is called an *inference* test. Here a paragraph is followed by seven or eight short

TABLE 85

Degree of Truth Test

Directions: No one knows just what the American people are thinking. There is need to find out just what convictions are most firmly held on some disputed issues.

Indicate your opinion about each of the statements on the following pages by drawing a circle around the one of the numbers in the margin which expresses your judgment. The meaning of each number is as follows:

Mark:

[✓ +2]	+1	0	−1	−2	If you feel the statement is utterly and unqualifiedly true, so that *no one who had a fairly good understanding of the subject could sincerely and honestly believe it false.*
+2	[✓ +1]	0	−1	−2	If you feel that it is *probably true* or true in large degree.
+2	+1	[✓ 0]	−1	−2	If you feel that it is quite undecided, an open question or one upon which you are not ready to express an opinion.
+2	+1	0	[✓ −1]	−2	If you feel that it is *probably false* or false in large degree.
+2	+1	0	−1	[✓ −2]	If you feel that the statement is utterly and unqualifiedly false, so that *no one who had a fairly good understanding of the subject could sincerely and honestly believe it true.*

Work rapidly, but do not fail to circle one figure in each line.

1.	+2	+1	0	−1	−2	The churches are more sympathetic with capital than with labor.
2.	+2	+1	0	−1	−2	Dancing is harmful to the morals of young people.
3.	+2	+1	0	−1	−2	Jesus was more interested in individual salvation than in social reconstruction.
4.	+2	+1	0	−1	−2	To have experienced business men, who have made a financial success in private enterprise, hold the public offices of the country would be better than the present personnel.
5.	+2	+1	0	−1	−2	The modern laxness in the observation of Sunday is, on the whole, harmful to the best interests of the people.
6.	+2	+1	0	−1	−2	Foreigners who work in our mines or factories should be paid on the basis of the same standard of living which we would set for American homes.

(Watson, 1925. By permission of the Bureau of Publications, Teachers College, Columbia University.)

statements. The student is asked to consider whether the conclusions follow from the paragraph, as illustrated in Table 86. Prejudice is supposed to be indicated when persons check items which do not logically follow from the para-

TABLE 86

Inference Test

Directions: Mere facts may mean different things to different people. It is often important to know just what people think certain facts mean. In the following pages you will find several statements of fact and, after each, some conclusions which some people would draw from them.

Put a check (√) in front of each conclusion that you believe is fairly based upon the fact as given here. *Do not assume anything else than the evidence given in the statement here,* with all its terms understood. You are not to consider whether the conclusions are right or true in themselves, but only whether they are rightly inferred from the facts given in the statement. You may check as many as you believe to be perfectly sure and certain. Do not check any merely probable inferences.

Example: 6,500 students recently attended a conference in which the questions of race relations and of possible attitudes toward war were discussed, these being the problems the students felt to be most vital today.

___ The students were all pacifists.
___ The students were all militarists.
___ The students came from all sections of the country.
___ Some students are interested in the treatment of Negroes and Japanese in this country.
___ Some students felt war was wrong.
_√_ The question of attitudes toward war is considered by many students to be important enough to be discussed.

I. Statistics show that, in the United States, of 100 men starting out at an age of 25, at the end of 40 years one will be wealthy and 54 will be dependent upon relatives or charity for support.

___ 1. The present social order cheats the many for the benefit of the few.
___ 2. The average young man, under present conditions, cannot count on being wealthy at the age of 65.
___ 3. Most men are shiftless, lazy, or extravagant; otherwise they would not need to be dependent.
___ 4. The one man is living upon luxuries ground out of the bones of the mass of common people.
___ 5. Some day the workers will rise in revolt.
___ 6. None of these conclusions can fairly be drawn.

(Watson, 1925. By permission of the Bureau of Publications, Teachers College, Columbia University.)

graph. The fourth part, called a *moral judgment* test, asks a student to indicate approval or disapproval of instances which are described in short paragraphs (see Table 87). The fifth part is called an *arguments* test. Here one is asked to distinguish between strong and weak arguments in support of a particular question. The last part is a generali-

TABLE 87

Moral Judgment Test

Directions: Most actual judgments of right and wrong have to be made in concrete instances. Mere general principles are not enough.

In the following pages you will find several instances upon which the moral judgment of individuals would differ. Read each carefully. You may assume each fact as stated. Then look at the alternatives suggested below it. Place a check (√) in front of the one with which you most fully agree. If you do not fully agree with any, check the one which comes nearest to expressing your opinion about the incident.

Example: A man stumbled into his house, drunk with bootleg whiskey. He smashed up some of the furniture and beat his wife and children. Then he stole some money from his small son's bank in order to buy more whiskey.

___ His action is worthy of approval.
___ The people who tolerated the sale of bootleg whiskey were in some degree responsible.
___ The occurrence is worthy neither of approval nor of disapproval. It is quite indifferent.
_√_ It would be desirable to prevent such a thing happening again, if possible, by establishing a better type of character in the man himself.

I. In 1793 the government of the United States recognized the young French republic, and President Washington received Genêt, the French ambassador. At the time Paris ran red with blood, the jails were full of the nobles who had been driven from power, and the government was in the control of a few high-handed dictators.

___ 1. Washington was right in recognizing this government.
___ 2. It made no difference whether recognition was extended or not.
___ 3. Washington was unwise. The government should not have been recognized under such circumstances.
___ 4. If it was a government which the majority of the French people really wanted, then it should have been recognized.

(Watson, 1925. By permission of the Bureau of Publications Teachers College, Columbia University.)

zation test which is very much like the second part. Watson also secured information concerning one's approximate wealth, schooling, vocational and religious experience, and membership in political parties and clubs. The method of scoring is to obtain the sum of all points of credit allotted for statements favoring one point of view. A general picture of a person's prejudice may be had from a gross score, and the separate scale scores give a profile showing graphically the strength of opinion along the twelve lines of economic, religious, and moral values shown in Figure 64.

The scoring standards were secured by submitting the items to small groups of judges including teachers, industrial leaders, and psychologists. No items were retained in a preliminary form which were not agreed upon by at least 75 per cent of the judges. In the final form, items were excluded unless they showed some plausibility by being chosen by several persons among a sample of 200. The odd-even reliability was approximately .96 for total scores. The separate parts showed a median odd-even reliability of .80, and a median correlation with the total scores of .63. Part II, Degree of Truth, showed the highest correlation, .94, with the total, and also with total scores of the remaining parts, .42. The first part, Cross Out Test, showed low correlations with the other parts, .11. Watson concluded that each part of the test was reliable enough to use in attitude measurement and valuable because of its special contribution.

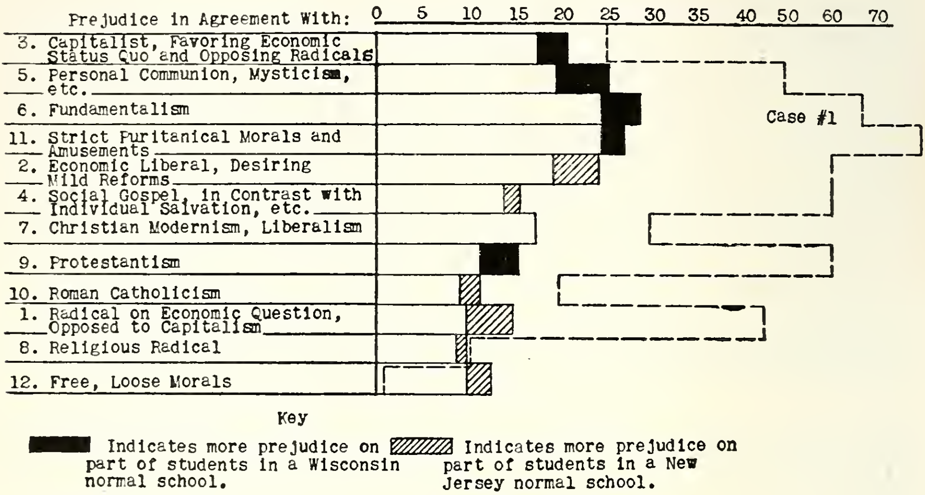


FIGURE 64. **A Test of Public Opinion**

(Watson, 1925. By permission of the Bureau of Publications, Teachers College, Columbia University.)

Correlations of intelligence and reading test scores with total prejudice were nearly zero. Part III, Inference, and Part IV, Moral Judgment, also show zero correlations with two group intelligence tests. Watson believed that the incorrect inferences were usually made not from failure to reason correctly among these college students, but from strong convictions which overcame intellectual conclusions.

Figure 64 shows the results of applying the Test of Fair Mindedness to two groups of normal school students, one in New Jersey and the other in Wisconsin. In general the scores are very similar, although the New Jersey students show consistently more preference for economic conservatism, religious mysticism, fundamentalism, and strict moral standards than the midwestern group. In other reports very marked differences have also been found among groups of persons who would be expected to show marked religious or economic preferences. Watson's Test of Public Opinion has in the past been used to appraise the effects of courses of instruction, but some of its items are now out of date.

B. Thurstone and Chave's Attitudes Scales

Another type of attitude scale was constructed by Thurstone and Chave (1929). These authors designed a single scale for attitudes toward an institution, such as the church. They first secured approximately 300 statements which showed a wide range of approval or disapproval of the institution in question. The method of selecting and scaling the items is discussed in Chapter 23. Table 7, page 36, shows the odd-numbered half of the items in their scale for measuring attitudes toward the church. Similar scales have been made up for evaluating attitudes toward a fairly large number of institutions and racial groups. The advantages of this type of scale are (*a*) that it insures a wide range of opinion among items, (*b*) that the items have been selected to give nearly equivalent steps in a scale, and (*c*) that items which are irrelevant or ambiguous are excluded. A person's score is simply the median scale value of the items which he checks.

C. Murphy and Likert's Study of Public Opinion

The work of these investigators was carried on over a period of five years among college students in several universities. They sought to determine what attitudes were held by these groups and what experiences had determined these attitudes. They used three sorts of appraisals of attitudes: self-rating inventories, reactions to picture and paragraphs, and autobiographies. They also secured measures of amount of specific information about controversial issues and records of scholastic success.

The self-rating inventories called attitude scales were constructed to evaluate opinions on:

1. Internationalism: war and peace
2. Imperialism: the use of force by great powers to maintain or extend their empires
3. Negro activities: social, economic, political, and educational

Some items for these scales were selected from questionnaires which had been extensively used. Others were designed to sample opinions in the three fields listed above. In the selection of items the following five criteria were applied:

1. All statements should be an expression of desired achievements not merely statements of fact, since persons with opposing attitudes often agree on questions of fact.
2. Clear statements involving only one issue should be used.
3. Each statement should be worded so that the modal reaction to it should be near the middle of possible responses.
4. Approval of about one half the statements and disapproval of the other half should express the same attitude.
5. The items in one scale should all measure the same attitude. This was achieved in part by requiring each item to show a high correlation with the total score, median .60.

Total scores were the sums of values assigned to various degrees of approval or disapproval of each statement. The retest reliabilities of the three scales when applied twice, with a three-week interval, were found to be approximately .85, using small samples of college students.

The median correlations between the scales were, internationalism with imperialism, .63; internationalism with negro activities, .40; and imperialism with negro activities, .34. A fourth scale on economic liberalism, which was not so widely used, dealt with organized labor and the distribution of wealth. Scores on this scale correlated .40 with internationalism, .39 with imperialism, and .30 with attitudes toward negro activities. The authors believe that their scales demonstrate "high generality rather than specificity in social attitudes." They also feel compelled to admit the presence of a general radicalism-conservatism factor as an explanation of the intercorrelations of these scales, and of data from autobiographies.

Films, photographs, and paragraphs depicting scenes of violence or force were also extensively used. Table 88 gives the questions that were asked after a picture had been shown of a wrecked automobile of a worker who refused to go out on strike. In addition to pictures of this sort, three short

TABLE 88

Appraisal of Attitudes by Photograph

(Questions asked after viewing photograph No. 14, showing a wrecked automobile of a workman named Langlois who refused to go out on strike)

1. Describe briefly in outline form your reaction to this photograph.
2. Indicate, by checking, how this picture affects you. Double check those which are especially intense.

   Excites
   Depresses
   Thrills
   Bores
   Interests
   Angers (irritates or enrages)
   Amuses
   Disgusts
   Others . . . . . .
   . . . .

3. In this situation, with whom do you sympathize?
4. What do you *like* or *dislike* in this photograph?
5. Why? (Answer briefly.)

(Total response to this photograph scored 2.)

(Murphy and Likert, 1938. By permission of Harper and Bros.)

motion picture films were used and a number of paragraphs. The results indicate that attitudes toward specific events could be measured reliably, but low correlations were found between items. This fact led to the conclusion that specific factors in each situation were large as compared with a general factor. Similar results were found for single items in the attitude scales.

The Bogardus (1925) Scale of Social Distance, shown in Table 89, was also applied by Murphy and Likert. It yielded specific scores of attitudes toward particular races and a general tolerance score whose odd-even reliability was found to be .88 among small groups of college students. This high coefficient was taken to indicate a high degree of generalization of attitudes toward persons in national and racial groups who are regarded as outsiders. Further analysis is needed to show the patterns of racial preference which will doubtless appear from a detailed study of separate items in such a scale. The total tolerance score correlated approximately .68 with the separate scores from the questionnaires on internationalism, attitudes toward the negro, and economic liberalism.

All of these opinion scores show only zero relationships with measures of specific information in the same fields or with intelligence test scores. On the other hand, a positive relationship between grades in college and radicalism scores on the opinion scales was discovered. Scholarship was more highly related to a radical stand for peace than to the other attitude scores. The authors believe that high scholarship usually involves much reading of recent literature, which happens to be largely radical in nature. The only other factor which corresponded with attitude scores was attitudes of parents. The authors conclude that the appraisals of attitudes dealing with public issues can best be studied from biographical material supplemented by systematic inventories. They suggest more thorough and longer biographies than are available at present.

TABLE 89

Bogardus' Test of Social Distance, Revised

	To close kinship by marriage	*To my club as personal chums*	*To my street as neighbors*	*To employment in my occupation in my country*	*To citizenship in my country*	*As visitors only to my country*	*Would exclude from my country*
Canadians	1	2	3	4	5	6	7
Chinese	1	2	3	4	5	6	7
English	1	2	3	4	5	6	7
French	1	2	3	4	5	6	7
Germans	1	2	3	4	5	6	7
Hindus	1	2	3	4	5	6	7
Indians (Amer.)	1	2	3	4	5	6	7
Irish	1	2	3	4	5	6	7
Italians	1	2	3	4	5	6	7
Japanese	1	2	3	4	5	6	7
Jews — German	1	2	3	4	5	6	7
Jews — Russian	1	2	3	4	5	6	7
Mexicans	1	2	3	4	5	6	7
Mulattoes	1	2	3	4	5	6	7
Negroes	1	2	3	4	5	6	7
Norwegians	1	2	3	4	5	6	7
Portuguese	1	2	3	4	5	6	7
Filipinos	1	2	3	4	5	6	7
Russians	1	2	3	4	5	6	7
Scotch	1	2	3	4	5	6	7
South Americans	1	2	3	4	5	6	7
Spanish	1	2	3	4	5	6	7
Turks	1	2	3	4	5	6	7
Roman Catholics	1	2	3	4	5	6	7
Fundamentalists	1	2	3	4	5	6	7
Liberal (Modernist) Protestants	1	2	3	4	5	6	7
Agnostics	1	2	3	4	5	6	7
Atheists	1	2	3	4	5	6	7
Trade Union Members	1	2	3	4	5	6	7
Socialists	1	2	3	4	5	6	7
Communists	1	2	3	4	5	6	7

Name..............................

Date..............................

(Murphy and Likert, 1938, p. 133. By permission of Harper and Bros.)

D. Attitude toward the Negro

Horowitz (1936) investigated attitude toward negroes by three sorts of tests which were administered to boys from the kindergarten through the eighth grade in various communities in or near New York City and in Georgia and Tennessee. Two of the tests used a page of 12 pictures of boys' faces, 4 of white and 8 of negro; all judged to be pleasant by a group of adults (see Figure 65). On one test, called "Ranks," the instructions were to "Pick out the one you like best, next best, next best . . ." and so on until they were all ranked. The score was the sum of the ranks assigned to the white boys' pictures; the smaller the score, the greater

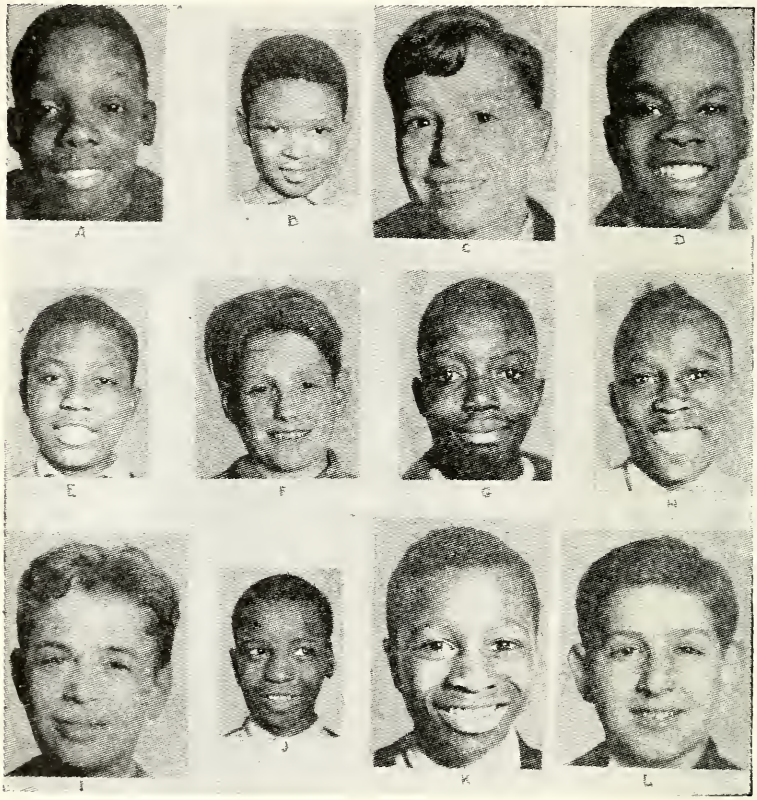


FIGURE 65. **Photographs of White and Negro Boys**

(Courtesy of Horowitz, 1936. By permission of the *Archives of Psychology*.)

the preference for whites. On another test, called "Show Me," the instructions were:

1. Show me all those that you want to sit next to you on a street car.
2. Show me all those that you want to be in your class at school.
3. Show me all those that you would play ball with.
4. Show me all those that you want to come to your party.
5. Show me all those that you want to be in your gang.
6. Show me all those that you want to go home with you for lunch.
7. Show me all those that you want to sit next to in the movies.
8. Show me all those that you would go swimming with.
9. Show me all those that you'd like to have for a cousin.
10. Show me all those that you want to be captain of the ball team.
11. Show me all those that you want to live next door to you.
12. Show me all those that you like.

The score was the per cent of total selection which were white boys. The higher the per cent, the greater was the preference for whites.

A third test consisted of 39 photographs showing activities of playing marbles, choosing sides for baseball, hand wrestling, sitting around outdoors, playing a piano, listening to a radio, playing checkers, eating in an ice cream parlor, eating dinner in a home, and common situations in lavatory, workshop, museum, library, and schoolroom. Each situation was posed twice, once by four white boys and again with a negro boy substituted for one of the whites. Each picture was observed in turn and the examiner asked, "Do you want to join in with them and do what they're doing along with them?" The scores were the difference between the number of times all-white groups and groups with one colored boy were chosen. The results of each test proved to be fairly reliable when checked for retest consistency. The general trends for New York City whites are shown in Figure 66. Here the ranking test shows the most prejudice, and this is true for all grades. The Show Me test indicates little prejudice at year 5 but there is a rapid increase until approximately year 8. The Social Situations test showed very little prejudice at year 5, and a very slow increase to year 14. Intercorrelations between these tests increased with age. The New York City group showed responses similar to the southern groups. The degree of prejudice, as shown by a tendency to exceed chance scores, was greater in the less specific test situations. Horowitz concludes that prejudice is a socially developed response which is derived from various sources. The most important source, at first, is probably the attitudes of friends and relatives. Later, the attitude may be the result of specific experiences.

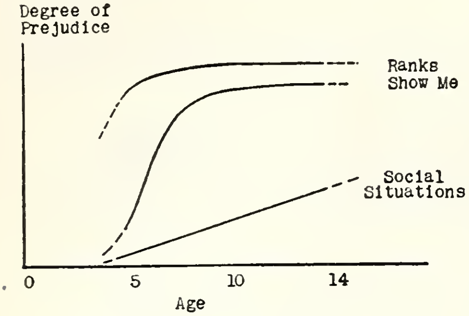


FIGURE 66

Growth of Attitudes toward the Negro

(Horowitz, 1936, p. 25. By permission of the Editor, *Archives of Psychology*.)

E. Reasons for Acts

Direct approaches are inadequate when the subjects do not wish to reveal their true attitudes. Brandt (1937) made an approach to the analysis of replies to questions of ethical significance which is interesting because it is somewhat indirect. In a group test situation the following instructions were read:

> In each of the situations given below, a person would have to make a choice between two possible actions. For each situation write down *all* of the points he should consider before he decides which to do. Remember, we are not in the least bit interested in which decision you think is the right one, or which one you would make. You are to concern yourself with writing down *only the things that are to be considered* in making the decision.

At this point the experimenter gave orally an example (losing one's fountain pen — to buy a new one or not). He then discussed with the group the particular things involved in a consideration of the situation presented. Samples of the situations used are:

I. Property.
   A. Stealing: Finding an unlocked car on the street; to drive it around awhile and return it or leave it alone?
   B. Destruction: Finding some things one wants in a magazine in the public library; to tear out some pages or not?

II. Persons.
   A. Friends: Finding out about someone who has been doing serious injury to other people; to report him or not?
   B. Group loyalty: Under what circumstances would one join a club?

III. Authority.
   A. Parent: Trying to get along with the family or leaving home
   B. Law: Obeying a law or school ruling even though you don't see the reason for doing it
   C. Religion: Going to Sunday movies or not

IV. Social.
   A. Conventions: Having to do something that injures your physical health in order to be one of a group
      Indulging in one's own tastes for clothes regardless of what others are wearing
   B. Superstitions: Finding a ladder in one's path; to go under it or avoid it?

The answers were classified into six groups on the basis of type of goal. These are illustrated in Table 90 for the item, "Indulging in one's own tastes for clothes regardless of what others are wearing." The boundaries of these classes are doubtless somewhat vague, but it seems probable that in a more extended and refined study a high degree of uniformity would be found among judges in classifying the answers.

TABLE 90

Bases of Judgments

For the item "To indulge in one's tastes for clothes regardless of what others are wearing."

1. Self-regard (17 per cent)
   Will I be looked down upon?
   Will people laugh at you?
2. Parental approval (8 per cent)
   Do my parents think them O.K.?
3. Friends' approval (7 per cent)
   What do my friends think?
4. General welfare (8 per cent)
   Wear the best you can.
   Clothes do not make the man.
5. Objective or practical considerations (39 per cent)
   Does the clothing fit? Are they appropriate for the weather? Where are you living? What is the comparative cost and durability.
6. Social institutions (21 per cent)
   Who are the people you associate with? Have I social ambitions? Do you like to dress like other people?

(After Brandt, 1937. By permission of the
University of Iowa, *Studies in Child Welfare.*)

The per cents in the parentheses in Table 90 are the distribution of all answers given by all persons to all the items. Brandt found that individuals varied a great deal from this average distribution of answers. Some subjects emphasized their self-regard, others parental approval, and others social orientation. This is valuable diagnostic technique which should be more widely used. It is interesting to note that moral or religious considerations were nearly lacking in this group. The answers would doubtless vary if the study were conducted in a Sunday school.

Responses to particular questions showed that there was a good deal of hedging or evasion or making of particular exceptions for unethical acts. Thus, most of the group were willing to take an unlocked car for a spin on such considerations as: "Is there sufficient gas in the tank? Can they drive?

Can they get away with it? What are the chances of wrecking the car? Is there an urgent need? Who is the owner? Will he care?" Considerations of this sort were most frequent on items which seemed to be or to have been of immediate concern to the subjects.

F. Allport and Vernon's Study of Values

The Study of Values Questionnaire is interesting in that it attempts to appraise attitudes in the six fields which were described by Spranger (1928). The test includes 120 items, distributed equally in the six fields. An inspection of the test shows the following sorts of items classified as:

1. Theoretical: Discovery of natural laws, mathematical relations, and scientific facts
2. Economic: Activity in real estate, finance, industrial development, and vocational training, and practical applications in general
3. Aesthetic: Indulging in artistic appreciation or composition in poetry, literature, music, dancing, architecture, and scenery
4. Social: Responding to others' needs with unselfishness and sympathy as shown by charity, freeing of slaves, hospital and social service work, and general friendliness and co-operation
5. Political: Managing governmental and legal affairs, curtailing charity, aggression in war, debating, playing on athletic teams, organizing, holding a seat in congress, exploration of little known parts of the world, and acquiring professional and social prestige
6. Religious: Abolishing war, laying up treasures in heaven, being reverent in a church, believing in God, comparing religious faiths, and evaluating life as a whole

The test has two parts. All the items in Part I require that a preference be expressed between two fields of activity, Table 91. These items combine judgments of *Yes* or *No* with ratings from 0 to 3, to show amounts of preference. The first item demands a judgment between discovery of scientific laws and applications of scientific laws; the second item, between aesthetic creation and social compatibility. The items in Part II require that four choices be rated in order of their appeal, Table 91. In Item I the first choice

TABLE 91

A Study of Values

Part I

Directions: A number of controversial statements or questions with two alternative answers are given below. Indicate your personal preferences by writing the appropriate figures in the right-hand columns, as indicated:

	(*a*)	(*b*)
If you agree with alternative (*a*) and disagree with (*b*), write 3 in the first column and 0 in the second column, thus	3	0
If you agree with (*b*); disagree with (*a*), write	0	3
If you have a slight preference for (*a*) over (*b*), write	2	1
If you have a slight preference for (*b*) over (*a*), write	1	2

Do not write any other combination of figures after any question except one of these four.

There is no time limit, but do not linger long over any one question or statement, and do not leave out any of the questions, unless you find it really impossible to make a decision.

	(*a*)	(*b*)
1. The main object of scientific research should be the discovery of pure truth rather than its practical applications. (*a*) Yes; (*b*) No.		
2. Do you think that it is justifiable for the greatest artists, such as Beethoven, Wagner, Byron, etc., to be selfish and negligent of the feelings of others? (*a*) Yes; (*b*) No.		

Part II

Directions: Each of the following situations or questions is followed by four possible attitudes or answers. Arrange these answers in the order of your personal preference from first to fourth by writing, in the left-hand margin,

1 beside the answer that appeals to you most,
2 beside the answer which is next most important to you,
3 beside the next, and
4 beside the answer that least represents your interest or preference.

You may think of answers which would be preferable from your point of view to any of those listed. It is necessary, however, that you make your selection from the alternatives presented, and arrange all four in order of their desirability, guessing when your preferences are not distinct. If you find it really impossible to guess your preference, you may omit the question.

1. Do you think that a good government should aim chiefly at —

.... *a*. More aid for the poor, sick, and old
.... *b*. The development of manufacturing and trade
.... *c*. Introducing more ethical principles into its Policies and diplomacy
.... *d*. Establishing a position of prestige and respect among nations

(Allport and Vernon, 1931. By permission of Houghton Mifflin Co.)

emphasizes social values; the second, economic; the third, religious; and the fourth, political. Parts I and II are combined into a final score for each field, which may be expressed in profiles, such as those in Figure 67. Here the mean scores

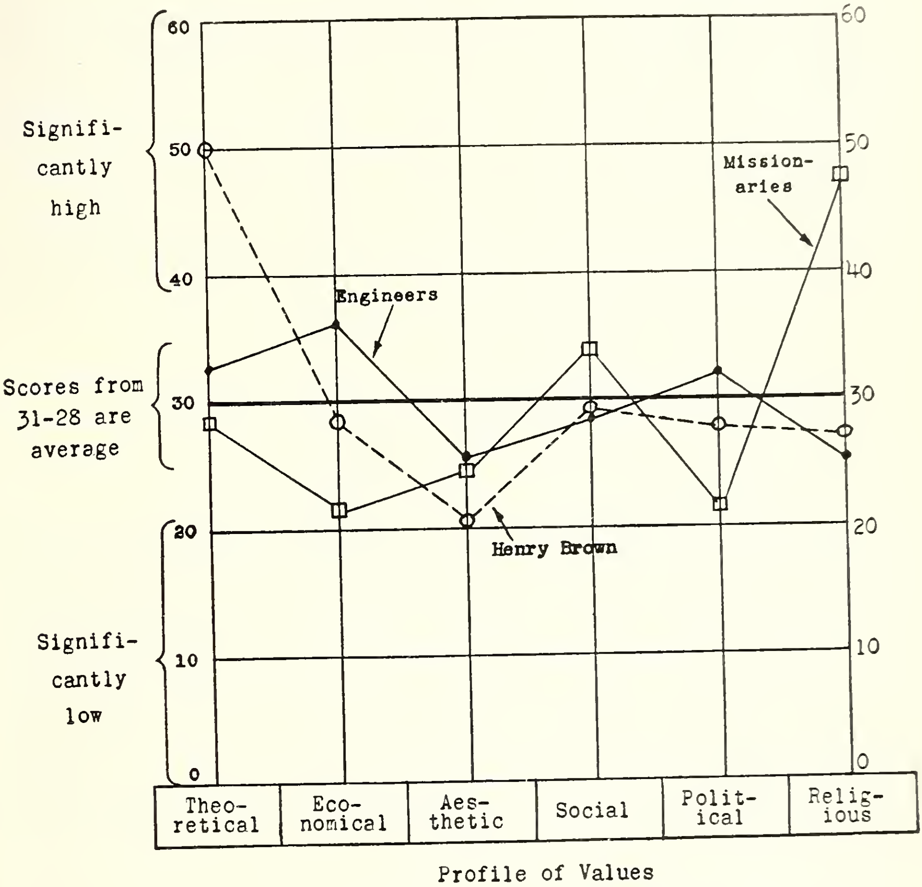


FIGURE 67. **Profile of Values**

Mean Scores of 64 Engineering Students and 80 Missionaries

(Arranged from Allport and Vernon, 1931. By permission of the *Journal of Abnormal and Social Psychology*.)

of a group of 61 engineering students show that their greatest interests lay in Theoretical and Economic activities, and their least in the Aesthetic and Religious. The highest interests for a group of 81 missionaries of both sexes are shown by this figure to be in Religious activities, the next highest in

Social, and the lowest in Political and Economic activities. A fairly large number of studies using this questionnaire have reported differences of this sort between various professional and social groups.

A statistical evaluation of this form has been reported by Cantril and Allport (1933), who reviewed fourteen published studies of this questionnaire. The original means were found to be correct for an additional group of 2,755 cases. The range of scores varies from a standard deviation of 5.5 in social values to 9.7 in economic. The repeat reliability is reported for 84 students over 100 days as follows:

Religious	.87
Aesthetic	.86
Economic	.79
Political	.76
Theoretical	.68
Social	.50

The authors point out that these figures show that the test of social values is less discriminating and less consistent than the other tests. The unsatisfactory results from the social values test are attributed to confusion in defining the term *social,* which led to the inclusion of several independent traits under this heading, and also to ambiguities on interpretation of some of the items. The authors question the existence of Spranger's "social type" and conclude (p. 272):

> On the theoretical side the evidence from recent applications of the Study of Values must be interpreted as establishing the Values, with the exception of social, as self-consistent, pervasive, enduring, and above all, generalized traits of personality. Several experiments demonstrate a clear relationship between values and conduct. They show that a person's activity is not determined exclusively by the stimulus of the moment, nor by a merely transient interest, nor by a specific attitude peculiar to each situation which he encounters. The experiments prove on the contrary that *general evaluative attitudes* enter into various common activities of everyday life, and in so doing help to account for the consistencies of personality.

According to Allport and Vernon, proof for the existence of these generalized attitudes lies in a number of findings, set forth as follows:

1. The internal consistencies of the various subtests: Items were selected which showed the highest relationships between item score and total subtest score.
2. The diversity of items in a subtest: The authors tried to select items which would test as many different applications of a generalized interest as possible.
3. Correlations between the Study of Values scores and activities, such as reading a newspaper, showing interest in clothes, securing high grades in college studies, attending church, speaking, handwriting, and artistic activities: Such correlations range from approximately +.25 to +.66 on small groups of students.

From an examination of these findings, however, the existence of these five general attitudes does not seem to the writer to be well established. The internal consistencies, which are generally low, would in any case merely show a coincidence of pattern in the persons tested. Such coincidences may occur from a number of factors rather than one of the five general attitudes selected. The moderate correlations between scores on the Study of Values Test and various activities might also be explained as resulting from a larger or smaller number of factors. Moreover, the test items for a single field appear to many to be quite independent. Thus, in the field of economics, there seems to be very little if any connection among interests in financial transactions, in vocational training programs, and in problems of practical engineering. In the field of politics, which is more accurately described as interest in *power,* there seem to be small similarities among such varied activities as debating, playing on athletic teams, curtailing charity, aggression in war, and exploring strange lands.

Furthermore, an inspection of test items reveals an overlapping among fields of interest, a fact which makes their identification difficult. For instance, the writer has found that many persons cannot distinguish among a search for truth in the *Theoretical* field, an attempt to evaluate life in the *Religious* field, and appreciation of true art in the *Aesthetic* field. There is also a marked similarity between unselfishness in the *Social* field and ethics in the *Religious.*

G. Factorial Analyses of Attitudes

The application of multiple-factor analyses to data of this sort may be of value in discovering underlying patterns. Such an analysis was made by Lurie (1937), using a 144-item test very similar to the Study of Values Test in content and construction. Six hundred students were given the test sheets. Only 203 returned them in complete enough form to be used, 128 men and 75 women. A seven-step rating scale, ranging from complete rejection to complete acceptance, was used for each item. Twenty-four scores were secured for each person by subdividing each of the six main divisions into four smaller divisions on the basis of (*a*) present interests, (*b*) ideals or standards, (*c*) preferences for associates or famous persons, and (*d*) beliefs or opinions. An analysis of these 24 scores by Thurstone's method resulted in the seven independent factors tentatively named:

1. *Social* or *Altruistic*
2. *Philistine:* aggressive, utilitarian, anti-cultural
3. *Theoretical:* interest in science, criticalness
4. *Religious:* doctrine and practice (not mystical)
5. *Open-minded:* liberal vs. conservative
6. *Practical:* pressure to do something
7. *Aesthetic:* superficial lip service to culture, negatively related to aesthetic ideals

Lurie states that the first four of these factors may be taken as stable categories which cover all six of Spranger's types. The Philistine factor is found to be large in both the Economic and the Political fields and also large in a negative way in the Aesthetic ideals. Reports by other investigators support these findings about the intercorrelations among scores on the Study of Values Tests. Lurie believes that the last three factors are temperament traits or adjustments which cut across the fields of interests, and which were not controlled in setting up the test. Lurie's findings also include some low correlations among scores which were supposed to measure similar interests.

The fact that Lurie found evidence for a basic unitary Social or Altruistic interest, whereas Allport and Vernon did not find it, needs an explanation. The evidence is not at hand for a complete explanation, but the different findings must have resulted from a different sampling either of interests or of persons. It may be that in Lurie's form a larger number of items clearly differentiate between altruistic and nonaltruistic activities than in the Study of Values Test. Or it may be that Lurie's subjects were actually more variable in altruistic tendencies than Allport's and Vernon's, and hence showed distinct individual differences in this field. A combination of both explanations may be nearer the truth. The results illustrate very nicely the point that the statistical factors are always related to the particular items used and the particular population tested. No experiment can conclusively prove the nonexistence of a factor.

Whisler (1934) designed a set of questions which were intended to evaluate "generalized attitudes." He avoided the use of specific situation items such as are found in nearly all the other inventories discussed, believing that a few questions such as those in Table 92 would yield as adequate a sample of attitude as a larger number of specific items. A factor analysis of 31 items was made, using 126 undergraduates. Six factors were reported and identified from the items which had the largest loadings (p. 292):

a. *Acceptance of conventional ethical principles*
b. *Enjoyment of momentary pleasure*
c. *Interest in conflicts and controversies*
d. *Desire to be an effective agent*
e. *Participation in casual social relations*
f. *Criticalness and interest in the truth*

TABLE 92

Questions on Personal Attitudes

Instructions: Take the words used as roughly indicating a three-fold scale into which responses to the question may be put. The scale refers to your judgment of yourself, not to what you think other people's judgment may be. Consider each question carefully, but if in doubt as to the proper checking, guess.

Abbreviations used: M, much or frequently; S, considerable or sometimes; L, little or rarely.

1. How much would your liking for an acquaintance of the same sex be affected by whether that person had a genuine liking for and interest in children, or was indifferent?

M ______________ S ______________ L ______________

2. How much enjoyment do you get out of doing or saying things which are quite shocking to most of the people who see or hear you?

M ______________ S ______________ L ______________

3. How much change has there been in the last two years in the type of people you prefer and seek to associate with?

M ______________ S ______________ L ______________

4. How much enjoyment do you get out of working with materials or making things with your hands?

M ______________ S ______________ L ______________

5. To what extent, in general, is your evaluation, judgment, and opinion of an acquaintance of the opposite sex based on speculation and thought as to how satisfactory the person might be as husband or wife?

M ______________ S ______________ L ______________

6. To what extent are you interested in politics? (That is, — local, provincial, national, or international politics)

M ______________ S ______________ L ______________

(*25 more items*)

(Whisler, 1934, p. 285. By permission of the Editor, *Journal of Educational Psychology.*)

The identification of these factors is not very conclusive, owing to the different interpretations which may be placed on an item. Thus, Item 16, "having a standard of goodness by which plays are judged," may be interpreted as (*a*) having a moral standard of goodness, or (*b*) the possession of any standard, or (*c*) referring to particular types of plays, or (*d*) referring to the rendition of a play rather than to its construction. Various interpretations of other items are probable. Less ambiguous statements would allow more accurate interpretations of the results. Yet Whisler's approach is an interesting one which has given results different from those of other workers.

Carlson (1934) reported the results of applying five of Thurstone's attitude scales to 215 seniors at the University of Chicago. The correlations between measures of intelligence by a group test and attitude scores were:

Prohibition	.036	Communism	.330
God	−.191	Birth Control	.211
Pacifism	.402		

These results lead to the conclusion that the more intelligent students are likely to be more liberal in attitudes toward these issues. Carlson also found that a multiple-factor analysis of these five-attitude scales and intelligence test scores yielded three factors. These were tentatively named intelligence, radicalism, and religiousness by an inspection of the various loadings.

The question of the existence of basic general attitudes, such as those named in the three studies just examined, will lead to useless controversies unless very careful definition and unbiased procedures are used. Apparently, if one is allowed to select some observations and ignore others, overwhelming evidence can be secured supporting almost any hypothesis about the nature of attitudes. Furthermore, test items are carefully chosen to represent a certain general attitude and if a random selection of persons is tested, that general attitude will usually seem to be supported by an analysis of the test results. In order to secure a complete picture of basic opinions and their relationships, items must be included which clearly represent all possible attitudes. The results of applying such a test to large normal populations will make possible a more adequate analysis of attitudes. Those factors described thus far may be found to be unique.

or they may be combined with other patterns or subdivided in more essential traits.

III. PRACTICAL RESULTS

The strength of various attitudes among students, employees, or voters is a major concern of educators, employers, advertisers, and statesmen. A large share of their time and effort is spent in trying to develop particular attitudes in their communities. Measuring instruments which may give accurate pictures of changes in attitudes are therefore in great demand and the field is rapidly developing. Some studies of nation-wide scope have been of little value because they neglected to secure measures of a representative group of people. Carefully controlled studies which measure the same persons twice are usually limited to a few highly selected students; hence their results are also of a very limited significance. The sampling of attitudes of large populations requires a great deal of careful planning and hard work. Sampling techniques are discussed in Chapter 5.

A. Modification of Attitudes

The effects of propaganda have been studied in a fragmentary fashion by a number of investigators. Generally, propaganda techniques include dissemination of reading matter, oral arguments, pictures, and films of actual scenes. Although many efforts at propaganda combine several of these techniques, several studies have tried to evaluate them separately. A few studies will be reported to illustrate the main results and some of the difficulties which have been encountered in attempting to measure changes in attitudes.

1. MODIFICATIONS DUE TO ACTIONS OR ARGUMENTS. Modifications of attitudes in line with the attitudes of instructors have been reported by Manske (1935) and Kroll (1934). Manske used Hinkley's scale of attitudes toward the negro. He reported that among 22 classes of high school students,

8 showed slight changes opposed to teachers' attitudes during 10 "non-indoctrinating" lessons taught by one of sixteen teachers. This study seems to indicate that teachers were able to be rather impartial in presenting material of this sort. Kroll measured the result of one semester's instruction in English history. He found that when the teacher held a radical point of view, scores on Harper's scales of social attitudes changed among 183 high school boys toward radicalism, but that when the teacher took a conservative position, the boys showed no reliable changes toward conservatism.

A number of studies report changes in attitudes due to purposeful instruction. Results from grade-school and college groups have been reported on attitudes toward racial groups, propaganda, treatment of criminals, patriotism, prohibition, war, and other topics. Murphy *et al.* (1937) wrote a thorough discussion of such studies, which may be summarized as follows:

1. High school groups are usually more susceptible to changes in attitude than college or adult groups but the changes may be less permanent.
2. Changes, generally in the direction of liberalism, usually accompany both high school and college instruction in social, economic, and political studies.
3. The attitudes of the instructor may be as important as the information which is discussed.
4. A violent argument or episode tends to make people take sides, for a neutral position becomes untenable. The negative response to the appeal may be as great or greater than the positive. The distribution of scores tends to become flattened or bimodal.

Biddle (1932) reported that among 350 high school and college students susceptibility to propaganda was greatly reduced by reading pamphlets on the techniques of propaganda. Knower (1935) found little difference between the effects of "emotional" and "rational" appeals concerning prohibition among 10,000 college students. He found that significant changes were larger among all groups when students read the appeal than when they heard it delivered orally. Other reports by Cherrington and Miller (1933) and Wilke (1934) indicate that speeches given in person or over the radio are at least as effective as printed presentations of the same material.

Changes in attitude which follow the viewing of a motion picture have been reported by Thurstone (1931) and his collaborators. For instance, a film which showed Chinese people as able and artistic caused a mean change toward more favorable scores on the scale of attitude toward Chinese equal to 16.9 times the PE of the change. Another film in which some Chinese were pictured in a more unfavorable light resulted in a mean change to a less favorable attitude equal to 2.2 PE. Another film which showed bootlegging as a vice resulted in no significant change in attitude toward bootlegging, which was already unfavorable. In other studies attitudes toward war, crime, and racial groups have, in general, showed changes in scores after viewing films, and these changes were retained for periods as long as 19 months with a gradual return toward a neutral position.

Murphy and Likert (1938) reported a study of the effects of written propaganda. Fourteen test items were selected which appraised attitudes toward war, imperialism, and negro activities. These were administered to two small groups of college students. A week later one group read propaganda material which argued for radical changes and the other group read material supporting conservative points of view. The material consisted of seven short excerpts from speeches by army officers, college professors, and persons well known in public life. These selections could be read in about thirty minutes.

A week after the propaganda material had been read, the test was repeated. Then each group was given material to read which supported the point of view opposed to that which had been previously presented. Four weeks later the attitude tests were repeated. These three tests showed how much each student was affected by this amount of conservative and radical propaganda. Although there was some shifting of opinion among students, there was no marked tendency for those with radical inclinations to vary more or less than other students. In one of the two groups there was a slight tendency for the more radical students to show greater susceptibility to propaganda. The lack of positive results may be due to several factors, such as the combining in one score of attitudes toward several different issues, the use of several short excerpts from speeches rather than one long concentrated speech on one topic, the delay of one to four weeks between reading the propaganda material and tests, and the sophistication of the students.

2. MODIFICATIONS DUE TO PRESTIGE. Marple (1933) reported that high school seniors, college seniors, and adults, in that order, were susceptible to the influence of majority opinion although the differences between groups were not very great. The opinions of these persons were secured on two occasions, the second a month later than the first. They were asked to mark 75 controversial statements about government, war, races, schools, and morals to show agreement, uncertainty, or disagreement. On the second occasion the group was supplied with a record of the answers which had actually been given on the first occasion. Changes between two trials in the direction of group opinion were for high school students 64 per cent, college students 55 per cent, and adults 40 per cent. A similar experiment showed that the influence of opinions given by a group of 40 experts was not quite so great as the influence of the opinions of one's own group although both were marked.

A study by Moore (1921) reported that conservative students accepted majority opinion more regularly than radical students, but Murphy and Likert (1938) reported no relationship between radical attitudes and changes in attitudes due to knowledge of majority opinion. They found that announcing the majority opinion orally after reading a statement on a second test did cause students to vary somewhat from the first test. However, the correlations between scores on the internationalism scale and shifts toward majority opinion were all nearly zero.

3. MODIFICATION CONCURRING WITH AGE. The work of Horowitz (1936) described above gives an excellent illustration of age-group differences in attitudes among grade-school pupils. Two other illustrations are given below which are typical of the many reports available.

Murphy and Likert (1938) reported that approximately five years after the first survey in 1929, a retest using three attitude scales was made of 129 individuals who had been graduated from college. The odd-even reliability of attitude scores was nearly the same as that reported for the first trial. The mean scores shifted slightly toward more liberal and radical attitudes. The most significant differences were found on the scales for imperialism and economic practices. The reasons for these shifts were discussed at length without any clear-cut conclusions. Data on personal incomes indicated that the shifts were probably not due to personal want. The authors believe that the change during five years probably reflects awareness of the seriousness of the causes of the widespread world depression.

The reasons for attending church were recorded by Kingsbury (1937) for a group of Protestant churchgoers in Chicago. The percentages of persons who checked eight reasons were found for the four age groups — 15 to 25, 26 to 35, 36 to 50, and over 50. Nearly 80 per cent of the youngest group checked "To formulate a philosophy of life," "To hear music and literature," and "To gain new friends," but only about 20 per cent of the oldest group checked these items. Thirty per cent or less of the youngest group reported that they attend church "to keep alive the spirit of Christ," "To encourage family attendance," or "from habit," whereas nearly 80 per cent of the oldest group checked these reasons. "To solve personal problems" was checked by approximately 50 per cent of all ages. "Just some place to go" dropped from 30 per cent to 7 per cent with 30 years of growth. "For reassurance of immortality" dropped from 30 per cent to nearly zero between the 20- and 30-year groups. This reason became increasingly important with advance in age, 30 per cent of the fifty-year-olds having checked it.

B. The Relation of Information to Attitudes

The question whether accurate knowledge about a situation accompanies a liberal attitude is one which can be answered by comparing information test scores with attitude scale scores.

A remarkably widespread study by Watson (1929) reported the application of a questionnaire on far eastern relations among 3,000 adults in church groups, prisons, business clubs, and schools. The questionnaire contains both information items and attitude items. The scores of information on Japanese issues correlated .82 with favorable attitudes toward the Japanese, and information on Chinese issues correlated .70 with favor toward Chinese nationalism. Another study by Manry (1927) found among college students a correlation of .69 between knowledge of international affairs and a favorable attitude toward "World Citizenship." Wrightstone (1934) found that historical knowledge correlated .58 with economic liberalism among 400 pupils in Grades IX to XII.

Reckless and Bringen (1933) found that among college students information about negroes and their problems had a mean correlation of .64 with favorable attitudes toward the negroes.

In contrast to these results are those of Murphy and Likert (1938), who found a zero correlation between very rough measures of information and liberalism in a study of internationalism and attitudes toward the negroes. Biddle (1931) reported correlations between zero and .26 between knowledge and unfavorable attitudes toward Filipinos. Japanese, and Chinese. Bolton (1935) found zero correlations for 700 college students between information and attitude toward negroes. The low relationships reported are probably due in part to sampling of narrow groups, to mixing of issues, and to the fragmentary nature of the tests.

The general conclusion is that well-informed persons usually take a fairly liberal or experimental view on controversial issues. Poorly informed persons are more likely to approve extreme radical or reactionary policies.

C. Comparison of Various Techniques

1. BIOGRAPHIES, INVENTORIES, AND GRAPHIC SCALES. The question, which techniques are most valuable in the study of attitudes, can only be answered in terms of their accuracy and their uses. The three most common techniques employ biographies, inventories, and ratings of general attitudes. Nearly all writers who have used case histories or autobiographies report that these yield the most complete accounts. A good deal of emphasis is therefore being placed upon securing accurate and complete histories. Comparisons of thorough case histories with measurement techniques have been made by several persons. Stouffer (1930) compared a scale of attitudes toward prohibition with a written account of activities, and with a graphic rating of attitude toward prohibition. Two hundred and thirty-eight college students first completed the scale, and then wrote accounts, using approximately 1,000 words, of their experiences with the prohibition law and drinking liquor. Four judges rated these accounts on a graphic scale of indulgence. The agreement of judges found by correlating the rating of each judge with the others ranged from .83 to .89, mean .87. These figures show a very high degree of consistency. The correlation between scores on the attitude scale and the composite rating by four judges was .81.

The same students also rated themselves on two occasions with a graphic rating scale for attitudes toward prohibition. When the two self-ratings were combined, it was found that the combined self-ratings correlated .80 with Smith's scale and .80 with judges' ratings of accounts. These correlations are all high enough to suggest that the three methods were appraising the same aspects of persons with about the same accuracy. It is probable that each of the three methods introduces unique aspects which may be of importance in particular evaluations.

Another study by Watson (1925) compared scores on his test of public opinion with short descriptions written by friends. The following account is a description of the person whose profile is shown in Figure 64. Both the profile and the account show extremely marked preference for fundamental Protestant beliefs and strict morals, and a less marked preference for mild economic reforms.

> This man is a farmer in a small Wyoming town. He is the leading pillar in the Methodist church. When I visited one Sunday he was drilling the children in the Sunday School on the names of the persons and the significant details of each of the Old Testament miracles. His neighbors consider him, next to _____________, the most stubborn man they have known. He is a rank-and-file member of the Farm Bureau, so should share some of their progressive economic ideas (from Watson, 1925, p. 46).

A much more elaborate study by Murphy and Likert (1938) also showed close correspondence between scores on their test and autobiographies written by students. Several cases were found, however, where apparent inconsistencies were cleared up only by a very careful rereading of the autobiographies.

Thurstone and Chave (1929) compared scores on the Attitude-toward-the-Church Scale with students' self-ratings made at the same time. A graphic rating scale was used which consisted of a horizontal line and the words "strongly favorable to the church" printed at one end, "neutral" in the middle, and at the other end, "strongly against the church." The correlation between scores and ratings was

.67, a fairly high figure. The same students also indicated whether or not they attended church frequently and were active members in a church. The frequent churchgoers were ten times as numerous among those whose scores were favorable to the church as among those whose scores were unfavorable.

2. MEDIAN SCALE VALUE VS. TOTAL SCORE. Likert (1932) applied several of Thurstone's scales using two scoring techniques. First, he used Thurstone's method of having a person simply check the statements with which he agreed, a procedure which yields a median-scale-value score. This score implies that a person agrees with all the statements which are less extreme than his own median statement. In practice this may not be the case. Likert also used a total score method which required a person to rate the same items on a five-step scale, using the words: (5) *strongly approve,* (4) *approve,* (3) *undecided,* (2) *disapprove,* (1) *strongly disapprove.* A person's score was secured by adding the points checked on all the items. A comparison of the results of these two techniques showed that the retest reliability for the median scale values was .76, whereas that for total scores was .85. The scale values correlated with total scores .88. Thus it appears that while both methods of scoring yield similar results, the total-score method shows considerably more self-consistency. The reasons advanced for this superiority of total scores are the fact that each item bears a share of the score directly, and that persons can probably express themselves more accurately with a five-step scale than with only two choices, as when given the choice of either agreeing or disagreeing. The total scoring technique can be applied without the elaborate scale construction which is preliminary to the median-scale-value scores. Thurstone's scale construction has the advantage, however, of guaranteeing the inclusion of items which represent a wide range of attitudes and an equal number of items at each level.

These comparisons of various techniques lead to five tentative conclusions:

1. A scored biography and a questionnaire yield, under fairly well-controlled conditions, about equal retest consistencies, ranging from .60 to .80 depending upon the range of scores in the group, and the ambiguity of the items. Scores which depend upon the summation of a large number of items show greater retest or odd-even reliability than scores which depend upon a few judgments, or on a median value.
2. Self-ratings and self-inventories of activities seem to have about the same validity, as shown by agreement with various criteria. The selection of satisfactory criteria is particularly difficult.
3. The graphic rating of single general attitudes, as toward prohibition, is likely to yield retest reliabilities in the neighborhood of .70 among college groups, using the optimum number of steps.
4. Convenience and economy in appraisal lies with the verbal inventories and graphic scales.
5. All of the direct appraisals of attitudes are subject to intentional misrepresentation which are hard to detect. No indirect or observational methods have been widely used.

D. Group Norms

A fairly large number of studies have been reported of differences in attitudes found in age and occupation, sex, race, locality, and educational and economic groups. Such differences have often been cited as indications of the validity of the scales. Most of these reports require a great deal of study because the interpretation of results involves the simultaneous evaluation of all factors which might contribute to attitudes. Thus a small difference between attitudes of men and women toward war may be due to differences in age, training, and occupations, rather than to the difference in sex. No studies have come to hand where all these factors have been held constant, although some studies have produced significant results. In view of the complexity of the material the reader is referred to the very able discussions by Murphy *et al.* (1937), Rundquist and Sletto (1936), and Allport (1937).

IV. NEEDED RESEARCH

Research in this field is considerably behind research in the field of interests, discussed in Chapter 15. This state of affairs is probably due to the intangible nature of attitudes and their frequent lack of immediate importance. Additional research is much needed outside of the classroom. Most of the studies reported were made in classrooms where detached attitudes are often found. Comprehensive studies of attitudes where age, race, experience, and other factors are evaluated are going to be of great significance. The reduction of misrepresentation is a pressing problem, the solution of which may come from indirect self-ratings, or observations made by others. Lastly, the construction of less ambiguous test items is feasible. With such items, thorough factorial analyses may indicate more clearly the basic patterns of attitudes of a group of persons.

17

MODES OF ADJUSTMENT

Free Association and Inventories

I. INTRODUCTION

Modes of adjustment in the broadest sense include the goals which have been discussed in the last two chapters, but in a narrower sense, used in this chapter, they may be restricted to the ways in which a person approaches a goal.

The earliest studies of adjustment included miscellaneous collections of items which were thought to be associated in some fashion. Today scales of this sort are still prominent, but they are widely criticized because they yield scores which have no clear interpretation. They are gradually being supplanted by more analytical appraisals. Numerous attempts, however, to classify modes of adjustment have not yet led to uniform and widely accepted categories. A good deal of intensive research is now going on to discover the basic or elemental factors in adjustment.

The procedure is about as follows: First, an observer reports specific reactions to specific situations. Then comparisons of these reports are made to show individual consistency and individual differences. If an individual consistently differs in some fashion from other members of a group, he is said to possess a *trait.* Traits were at one time thought of as disconnected classes of behavior, such as honesty and dishonesty, but all recent studies have shown that measures of traits result in nearly normal frequency curves, or, at least, continuous frequency distributions. Traits are now considered to be continuous in the sense that they usually vary from small to great amounts in a group of persons, without any large gaps which would show a sharp division, as for instance, between honesty and dishonesty.

If a number of individuals show similar combinations of traits, they are said to belong to a *type.* Types in this sense are sometimes called *temperaments* or *syndromes,* and are illustrated by such concepts as melancholia, introversion, inferiority, and leadership. No sharp line can be drawn between traits and types.

Observed combinations of traits have led to theories concerning the nature of mental and physical organization. Nearly all such theories state that a living person has a group of basic physical needs which create tensions or incentives, such as desires for food, water, or exercise at various times. Tensions vary, depending upon variations in physical structure. This fact has given rise to a number of studies of bodily types and of corresponding personality traits. A recent approach to the measurement of varieties of human physique is given by Sheldon (1940). Furthermore, tensions drive the person to activity and to acquiring new ways of relieving them. Learning results in cultural ideals which may become very powerful, and which often conflict with physical tensions and also with one another. Under the general heading of traits are found three aspects of behavior:

1. *Variations in energy:* amount, endurance, and distribution in various directions
2. *Ways of working with people:* tendencies to use or serve other persons, cultural ideals (often called character traits)

3. *Integration or disorganization:* moods or procedures, and amounts of success in resolving various conflicts (Procedures are sometimes called psychological mechanisms.[1])

These three classes of traits are further subdivided by many writers, and the boundary lines among them are not very clear. It is of course important to secure the most adequate explanation for a person's modes of adjustment, but before explanations are attempted, one should be very sure about the facts. We shall, therefore, turn first to examination methods, and later to the theories of traits. Five methods which are commonly used will be illustrated. These are called free association, questionnaires, conduct measures, observation, and laboratory methods. The first two methods will be described in this chapter and the last three methods in the next chapter. An evaluation of all five methods is given in Chapter 19.

II. FREE ASSOCIATION

Free association methods allow a person to respond to a stimulus without any assigned limit to his responses. Practically, there are always intellectual limits or social taboos, but the subject is asked to respond without inhibition. The method was used by Galton in 1869, who read a list of words and asked the subject to respond after each word with the first idea that came to mind. The method was also

---

[1] These include such terms as:

a. Rationalization: relieving of a guilty feeling by attending to and believing only those arguments which justify the behavior

b. Projection: relieving a tension by disclaiming personal responsibility for one's acts

c. Hysterical traits: relieving of a mental conflict by violent motor activity, paralysis, or by marked sensory dysfunction and anesthesia

d. Repression: relieving a tension by splitting off from consciousness an unpleasant idea or fear or unsocial desire

e. Fixation: relieving of tensions by identifying oneself as belonging with someone else, or in extreme cases being someone else applied by Jung (1916) to detect memories of painful experiences which a person could not readily recall. Jung used a list of 100 words and noted unusual responses, such as stuttering, repetitious, queer mannerisms, and very long reaction times. The time between stimulus and response was recorded with a stop watch. Another application of a free association method was made by Rorschach (1921), who used a series of ink blots as stimuli. The ink blots have been much more widely used than the word lists, apparently for two reasons. One is that stimulus words have special connotation in various parts of the country and in various countries so that the results of one locality are not comparable to those of another. The other reason seems to be that the interpretation of results of ink blots has been much more thorough and meaningful than that of the results of verbal responses.

A. The Rorschach Test

In 1921 a work by Rorschach, a Swiss psychiatrist, was published which described an elaborate method of determination of modes of behavior from a person's verbal responses to ten ink blots. Five of the ink blots consist of various shades of gray, as shown in Figure 8, page 32. Two of the blots are gray with one shade of red, and three are entirely in color or nearly so. Various shades of red, yellow, green, orange, and blue are used. Today there are several hundred technical reports on the use and interpretation of these ink blots, and a research exchange for publications concerning this technique has been established.

1. Administration. The administration of this test should take place in a quiet room where the subject is placed so that he may give his whole attention to the stimulus card. The directions from Beck (1932), p. 33, are as follows:

> "What could that be?" or "What do you see?" The cards must always be presented in the same order, and when first presented

must always be in the same position for every subject. The subject is permitted to turn a card as he pleases, but he is not allowed to look at it from any greater distance than his own arm's length. It stands to reason that he is not permitted to see any of the cards in advance of the test. Examiner notes the responses verbatim. If the subject is unresponsive, or shy, it is permissible to encourage him to the extent of saying "But most people see something," or if subject tends to give cards up after one response, he may be encouraged by saying to him, "Most people see more than one thing." Rorschach adds that urging to the point of getting at least one response for each card is permissible. Nothing that in any way suggests any actual response may be said to or asked of the subject, even in answer to his own questions. No time limit for any card or for the test as a whole is set, and the subject is permitted to respond until he gives the card up of his own accord.

Although Rorschach and Beck have stated that reaction times are not significant in interpreting results, some investigators record them and believe them to be useful.

2. Scoring and Interpretation. The scoring of the results with this test takes place along four lines:

1. *Total number of responses*
2. *Apperception mode* (This is an indication of whether the subject perceives the ink blot as a whole or as isolated parts.)
3. *Qualities perceived,* such as form, color, and movement
4. *Content of responses* (This refers to the specific objects which are recalled, such as animals, trees, landscape, and people. Content also includes an account of the number of common responses and original responses.)

The interpretation of scores follows in general the logic and experience of Rorschach and a number of other psychiatrists. The total number of responses is somewhat associated with intelligence and verbal fluency. However, the total number of responses of a particular kind is not considered to be so important as the proportion of responses. Thus Beck (1937, p. 194) , who applied and collected large amounts of Rorschach data, found that the normal individual generally produced 31 responses of which 6 were reactions to an ink blot as a whole (W), 21 to a frequently selected detail (D), and 4 to a rarely selected detail (Dr). A deviation from this usual distribution toward a larger proportion of responses to the whole is interpreted as indicating a tendency to "make a broad survey of presented material: the strategic approach," in normal adults. An extreme deviation of this sort would indicate a tendency toward expansive generalities and a neglect of obvious details. Conversely, a deviation toward a larger proportion of response to details is interpreted to mean a habitual attention to the concrete, and a more practical approach. Extreme deviations of this sort, when rare details are frequently observed, may indicate a precise or minute discrimination tendency, and in abnormal cases pedantry and obsessions of thoroughness and cautiousness. The diagnosis of impracticality and lack of intellectual organization is often indicated. Feeble-minded persons are often found to lack the power to perceive ink blots of this sort as meaningful patterns. They usually report associations of common objects with fragments of the ink blots. Apperception mode also reflects training and schizoid tendencies.

The interpretation of the qualities reported by the subject is also interesting. Imagined movements, when reported more frequently than usual, suggest a capacity of invention which in extreme cases indicate fantasies or delusions and mental conflicts. When color qualities are reported more frequently than usual, this is taken to mean habitual impulsiveness, egocentricity, capacity for intense emotional experience, and, in extreme cases, violence and flightiness. An unusual number of clear responses to form indicate intellectual steadiness or introversion. Descriptions where form dominates but shading is noticed or where form dominates and color is mentioned are typical of well-adjusted persons with affective stability. Descriptions where color dominates and form is a minor factor are typical of egocentric organization. Where shading, called chiaroscuro, is very dominantly reported, persons are often markedly repressed.

The interpretation of specific memory content must take into account the background of the examinee. Thus, an unusually large number of references to human anatomy might be normal for a surgeon, but it might be indicative of a morbid tendency in a baker or a butcher. Oft repeated responses may indicate stereotypy, as in feeble-mindedness, or a special interest, or a recurring fear. The number of common and uncommon responses is indicative of either originality or of unusual usage of usual words. Breadth of interests is roughly indicated by the number of different fields represented in reports.

Beck emphasizes the great importance of interpreting the responses of a person in terms of his whole personality picture. Hence, no rigid interpretation of one element is possible; rather, it may indicate the direction of a force and the probable balance between trends. Beck gives complete records of 58 persons who are typical of seven classes of behavior: healthy adults, feeble-minded, depressed, hypomanic, schizophrenics, neurotics, conduct disorders in adults, problem children, and mental hygiene cases.

3. Reliability. In spite of the many reports of the Rorschach test, little can be said of its self-consistency. This situation is due to the fact that it has been developed as a clinical tool and used in clinics where opportunities and desire for retests were often lacking. Two studies report split-half correlations for various scores. Vernon (1933) found that such correlations were low for all scores and concluded that the test was nearly worthless statistically because of subjective scoring errors and the paucity of items. As would be expected, he also found low intercorrelations between various scores. On the other hand, Hertz (1934), using more definite norms and a carefully chosen scoring method, found split-half reliabilities from .60 to .90 and indicated that these could be raised still further by a refinement of technique.

Since the total number of responses of persons may vary, comparable scores cannot be secured for a particular quality, such as form, simply by adding up the responses which refer to form. Hence, indices are calculated to show how much a person tends to vary from a normal ratio of qualities.

It is, therefore, desirable to refer to norms in order to interpret the results.

4. NORMS. The establishment of quantitative norms for this test has been opposed by many persons including Rorschach himself, who believed that all results were related to each other and that items had little significance when removed from the total picture. Most of the norms which have been published, therefore, are lists of typical syndromes for various groups. They are usually based on only a few cases and lack any quantitative evaluation. In America, however, the extensive use of the test has produced a marked demand for norms. Norms on this test have not yet been secured for representative samples of all ages, but Hertz (1936) published norms for 300 adolescent subjects, and Gardner (1936) published norms for 100 normal adults. Beck (1937) criticized the extreme objectivity in scoring techniques so he published norms, without, however, clearly indicating the samples of the population to which they referred. He reported 328 subjects in all, who included 52 healthy adults of superior intelligence. The need for more accurate norms has been stressed by Sicha (1937), who submitted reports of 300 persons on the Rorschach test to four different scorers. The percentage of agreement ran from 82 per cent to 70 per cent. This indicated a considerable disagreement among the experts. Hertz (1938) has compared her scoring keys with those of Beck and Klopfer-Davidson (1937). The latter norms are set up from the experience of the experimenters. Klopfer believes that it is necessary to make a qualitative definition of certain Rorschach categories before attempting statistical treatment of the results. Hertz reports high agreement in the scoring

of cards 2, 6, 7, 9, and 10, and a very low agreement for cards 1 and 5. Further agreement may be expected when tables are available for age and race groups.

5. Validity. The comparison of Rorschach scores with other types of evaluation has as yet produced few significant results from a quantitative point of view. This does not, however, detract from the value of the test in the minds of most of its users. Rorschach claimed that his test distinguished among levels of intelligence, and particularly between feeble-minded and normal persons. Beck (1932) and Hertz (1934) reported that Rorschach's indicator of feeble-mindedness was of little value, but that other Rorschach scores not only diagnosed feeble-mindedness but also revealed peculiarities of the group. Beck found a correlation of .64 between Stanford-Binet scores and clarity of form on the Rorschach test. Hertz found that clarity of form correlated .46 with the Otis Self-Administering Test of Intelligence. Other Rorschach scores were less significantly related to ability. The relation of Rorschach scores to nonintellectual factors has been discussed in nearly 100 articles, which are ably summarized by Hertz (1935). A large number of investigators testify to the efficacy of Rorschach's methods in clinical diagnosis. Disorders such as schizophrenia, manic-depressive, post-encephalitis, and epilepsy have been diagnosed correctly from test results alone. Moral defects, inhibitions, neuroticism, negativism, feelings of inferiority, malingering, and expressive fantasy have been diagnosed from test results and confirmed by case studies.

6. Typical Reports. Of the many applications of the Rorschach material, three will be reported: one contrasts college and psychotic groups, another contrasts European and Moroccan adults, and a third reports a feeble-minded group.

Rorschach scores of 43 psychotic patients were compared with those of 52 college students by Vaughn and Krug

(1938). The psychotic patients included 38 women and 12 men, most of whom showed paranoid symptoms. The college group included 38 women and 14 men. The answers were recorded on mimeographed copies of the ink blots. The results show that the college groups had a mean total of 81 descriptions, and the psychotic, 57. All of the comparisons of various scores showed large similarities between the psychotic and college groups, but certain fairly well-marked tendencies appeared. The college group showed greater tendencies to report only color or only form or shading. The college group exceeded the psychotic group in reporting specific detail, open spaces, movements, and original responses. The two groups showed no significant differences in descriptions of animals, common situations, and popular expressions. The authors claim that this study indicates differences between psychotic and normal persons primarily in modes of perception of color and form, and that these differences are large enough to indicate the presence of serious disorder in one group of subjects.

The same investigators also compared the results of Bernreuter's Personality Inventory with Rorschach scores. They found a correlation of −.52 between Bernreuter's neurotic tendency and Rorschach's emotional stability, and a correlation of .78 between scores for introversion by the two methods.

A comparison by the Bleulers (1935) of Rorschach test results from 29 Moroccan peasants and from typical European adults brought out some striking trends. The Moroccans showed a marked lack of integrated responses. They interpreted the minute details in a fantastic manner, irrespective of the main parts of the ink blots. Abstractive generalizations, which are common in Europeans, were lacking in the Moroccans. On the other hand, the Moroccans and the Europeans showed very similar qualitative tendencies in reporting shape alone, color alone, kinesthetic impressions, accuracy of observation, animal responses, and reaction times. The Bleulers substantiated these results by depicting typical peasant reactions (p. 113):

> We find in the Moroccan a lively, extroverted interest in the episodes of the life around him, a marked enthusiasm under the influence of momentary events, a strong responsiveness to environmental stimuli; but in the long run and in the face of the great and important problems of life the Moroccan is a schizoid dreamer. He lacks the systematic, energetic, and persevering striving after outward success. In the long run he is not guided by the delight in earthly pleasures or by the sensitiveness to real suffering, but rather by a fantastic, egocentric, unreal, and strange attitude. He has moments of primitive, childlike impressionability, but in the broad lines of his development he is self-centered, introverted, rather independent of the outer world, and tends to prefer seclusion. It seems to us that this description of the Moroccans' emotional life, as inferred from the Rorschach Test, strikes the core of the Moroccan national character.

Beck (1932) reported a study of 69 feeble-minded whose ages ranged from 5 years and 10 months, to 65 years. Two-thirds of the group were less than 16 years of age, and their mean age was 10.8 years. The mean mental age from a Stanford-Binet test was 6.7 years, and the mean IQ, 55.1. The average interval between administration of the Binet and the Rorschach was 10.8 months and it ranged from 1 month to 42 months. An estimated MA was made for each person on the date of the Rorschach test, assuming the IQ to have been constant. These estimated MA's were correlated with various Rorschach findings as shown in Table 93.

This table shows a tendency for the higher scores on the Stanford-Binet to be associated considerably better than chance with the number of whole responses, the number of movement responses, and per cent of accurate clear form responses. There is a slight negative relation with the per cent of original responses, which seems to be due in this case to the fact that the idiots would give responses which were apparently not related to the test situation. The other correlations indicate a chance coincidence.

TABLE 93

Correlations of Binet Mental Ages and Rorschach Test Results

69 Feebleminded Subjects

	r	PE_r
1. Number of whole responses (W)	.474	.06
2. Per cent of clear form responses (F +)	.640	.05
3. Number of movement responses (M)	.256	.08
4. Number of color responses (C)	.012	.08
5. Per cent of original responses (O)	.391	.07
6. Total number of responses (T)	.012	.08
7. Number of responses to rare detail (Dr)	.148	.08
8. Number of responses to detail unrelated (Do)	.069	.08

(After Beck, 1932, pp. 71–74. By permission of the *Archives of Psychology*.)

Beck points out a number of uncontrolled factors which would tend to lower whatever correlation there might be, such as the ambition to make a good impression, verbalization, and the small number of cases. He claims that the Rorschach method as an instrument for measuring intelligence has the advantages of being relatively free from the effects of schooling at all ages, and from the emotional tensions set up by most test situations.

The main contribution of the Rorschach technique is the attempt to get a more complete picture of the way in which a person attacks a problem of organizing rather vague pictorial patterns. The results are far from being a complete and accurate picture of one's range of information or use of information in mental work, but the method is a good one. The test gives results which are more analytical of processes than the achievement type tests.

The Rorschach method has some disadvantages. It depends, perhaps too much, on temporary moods. The interrelationships of qualities and content are so complex and intangible that conflicting interpretations are likely, even from well-trained examiners. The test situation is rather artificial or unusual. Up to date, not enough cases have been retested to show typical personal variations. The method, however, seems to be a promising one for the analysis of modes of adjustment.

B. Picture Interpretation and Composition

Another promising technique which has been little used is the interpretation of pictures. Morgan and Murray (1935) asked students to make up short stories about pictures in which persons of their own age and sex were portrayed in a variety of incidents. It was found that the stories disclosed the needs, fears, wishes, and emotional life of the tellers in a remarkable manner. Good co-operation and lack of self-consciousness are usual in this procedure.

Vance and McCall (1934) devised a set of pictures of children at play for measuring children's interests by the paired-comparison method, and Giles (1936) used pictures of men performing various kinds of work for the same purpose. Schwartz (1931) designed a set of 8 pictures portraying various social situations considered to be important for initiating psychiatric interviews with boys between 7 and 15 years of age. Later he brought out a set of pictures of social situations for use in interviewing girls. The responses of each interviewee showed both his insight and how much he identified himself with the subject of the picture.

A number of studies have reported attempts to find indicators of modes of adjustment, particularly psychotic tendencies, in freehand drawings and in copying of drawings. Quite a few psychiatrists use this method regularly to aid in discovering fantasies, delusions, unconscious mechanisms, and various aspects of personality. Guttmann and Mackay (1937) pointed out that drawings often illustrate hallucinations or neurotic traits in schizophrenic patients. The main features which characterize drawings of these patients are mixtures of letters, lines, fragments of objects frequently repeated in the same drawing, great detail, and lack of realistic representation. Large numbers of the details are unconsciously symbolic of sex and religious motives.

Kerr (1937) found that neurotic tendencies in children led to drawing long thin houses rather than houses of more usual dimensions. Other emotional disturbances apparently affected the drawing, but did not show any statistical significance.

Horowitz and Murphy (1938) have given an interesting description of several possible uses of toys and pictures in the revelation of modes of adjustment among young children. Relatively structured materials, such as dolls, pictures, and many sorts of toys, can be used to elicit reactions to conventional patterns. Materials which are relatively unstructured, such as clay, finger paints, dough, mud, sand, and ink blots, invite spontaneous expressions because, as they state (p. 134):

> Conflict or guilty feelings regarding messiness are apt to be brought out by dough or mud; repressed aggressive impulses may be stimulated to expression by offering balloons to be exploded, cars with sparking guns, rubber knives, scissors with material to cut, or block structures to be knocked down. Emotional trends toward melancholy or gaiety may appear in preferences for or resistances to music of different kinds and for colors of emotional value to the child. Preoccupations with food, toilet functions, or sex may appear in "free play" with household toys while inhibitions or conflicts regarding people may appear in the avoidance of dolls. Free movement among a variety of materials and functional creative use of them in contrast to stereotyped routine uses may indicate basic spontaneity or inhibition; these basic patterns of independence of reliance upon authority frequently determine the child's response to different types of "progressive" or "conservative" school treatment and are important to analyze in formulating any educational plan for a child.

The methods of presentation may vary from directions, such as, "Do what you like," to instructions which limit a person to choosing between pictures to show which illustrate more of some quality. The more controlled procedure is described as follows (p. 137):

In the specific study for which it was designed, the investigator attempted to measure the extent to which adult standards in relation to various material aspects of living had been accepted by the growing children. There is no reason to believe that the same kind of test could not be developed to tap subjectively developed values as well as values interiorized from the social milieu. The items for this particular form of the test were chosen as possibly representative of important aspects of living in the culture. They included landscapes, houses, dishes, clothing, sport, dancing, flags, cities, kinds of work for men and women. One page of a large scrap book was prepared for each item. On this page were pasted pictures giving several possible choices. Each child was asked "Which is the best . . . ?" The questions included such items as: "The best place to live (landscapes)"; "the best houses to live in"; "the best kind of dishes to use"; "the best way for a boy to dress"; "the best way for a girl to dress"; "the best work for a man to do"; "the best work for a woman to do"; "the best sport"; "the best city to live in"; "the best way to dance"; "the best looking flag." Inasmuch as any of the choices presented on any single page might have been adjudged "best" according to different frames of reference, in order to make the choice, the child had to act in accordance with a particular kind of orientation, functioning subjectively. After the first choice had been indicated, the next four choices were given in the order in which they were preferred. The subject placed cards over his choices as they occurred in order to eliminate distraction and confusion. This kind of ranking was thought to make possible scoring on a more subtle basis than a single "best." As each group of pictures contained some representation of the ascertained familiar, the force of familiarity in determining the value judgments of the children could be measured. For example, in the set depicting cities, one picture was noticeably like the county seat to which these children made weekly trips. The range of houses included a cabin and a palace, but offered several which resemble closely types of homes indigenous to the community.

The clearest indication of the accretion of culture values came out in responses on a flag item. The page presented for inspection contained twenty-four different flags. The children were asked, as in the other items, to rank their choices in order of preference in answer to the question, "Which is the best looking flag?" The youngest children tended to choose the Siamese flag, with its white elephant on a red background. As the older children were tested, another trend was noticeable. They tended to choose the American flag, or flags containing elements similar to those that make up the American flag, stars, stripes, and the three colors. Finally, the American flag was almost the invariable first choice, and all the other rankings were given to flags bearing some resemblance to the flag of the United States.

These methods of investigation are less time consuming than the time-sampling or continuous-observation methods, and less inhibiting than direct-question methods. The use of pictures and materials to appraise behavior will undoubtedly be greatly refined and extended to older age groups in the near future.

C. Word Association Tests

Following the technique of Jung, Kent and Rosanoff (1910) secured the first response to one hundred common words from 1,000 normal persons (see Table 5). The subjects were simply instructed to give, after a stimulus word, the first word which came to them. The frequency of each response to each word was secured, and an individual's score was calculated to show the proportion of unusual responses. They found that in the average normal person 7 per cent of the total responses were unusual, and that 247 mental patients gave an average of 27 per cent unusual responses. The list of Kent and Rosanoff has been widely used on groups of college students, feeble-minded persons, negroes, and school children, so that per cents of unusual responses are available for these groups. However, the variety of explanations of unusual responses makes the interpretation of these results very difficult. Rosanoff (1920) and O'Connor (1934), working with what they thought to be fairly average adult samples, believe that unusual responses were largely due to unusual modes of behavior, such as autistic thinking and a tendency toward extroversion or introversion. McClatchy (1928), and several others, advanced the theory that persons with a large number of unusual responses are generally those with a large vocabulary or high intelligence. This is certainly true of college and high-school groups. Wheat (1931) found that those with low intelligence also gave a larger proportion of unusual responses than the average. This was due to lack of understanding of the stimulus word or lack of ability to vary responses.

Another standardization of results of a free association test was described by Wyman (1925). On the basis of teachers' ratings of interest in "intellectual, social, or activities programs," three groups of children were selected. The word association responses of each group were tabulated and compared. Three keys were then devised to allow three scores, one for similarity to those with intellectual interest, another to those with social interests, and a third to those with activity interests. A more elaborate study of a similar sort was reported by Kelley and Krey (1934). In this case, both teachers' and pupils' ratings were used to select groups of children who would be characterized by "courtesy, fair play, honesty, loyalty to fellows, mastery, poise, regard for property rights, and school drive." Word association tests were applied, and keys constructed to give scores in these eight divisions.

On subsequent applications, both Wyman's and Kelley's tests proved to have very little discrimination. Other groups of children, when rated and tested, showed low correlations between ratings and test scores. Kelley applied Hotelling's technique to the analysis of tests and ratings and found two main components which seemed to him to correspond to general social conformity and to assertiveness.

These tests of free association have in general been disappointing, but the basic assumptions and techniques are probably very worthwhile. The failure to be discriminating is probably due to the shortness or fragmentary nature of the answers. One is often left in doubt as to the nature of response when only the first word is considered. If a large number of responses were allowed for each stimulus word, as is the case in appraising the Rorschach ink blots, the classification of responses would be much more accurate.

Meltzer (1935) investigated children's associations with parents by encouraging the children to think aloud after each stimulus word. In the list of stimuli were the words *father* and *mother*. Unlimited responses to these two words were classified for pleasantness, degree of attachment, level of socialization, and other parent-child relationships. He reported 82 per cent agreement on the classification of responses by several investigators.

III. QUESTIONNAIRES

One of the earliest questionnaires is the Personal Data Sheet of Woodworth (1917). This is a self-appraisal on 116 items designed for a quick survey of neurotic symptoms among soldiers. The items were derived from psychiatrists' descriptions of neurotic and pre-psychotic patients. Mathews (1923) adapted this material for use with children, and Laird (1925) produced a form which has descriptive rating scales for each item. Thurstone (1930) published an inventory of 223 items from Woodworth, Laird, and other sources which is now widely used in America. Table 94 shows a sample of typical items selected by Mosier (1937) because of their discrimination value.

All of these inventories are scored to yield a general index of psychoneurotic tendencies. A general index, however, fails to show the particular difficulties which are present, and hence fails to indicate the most reasonable remedial procedures. To meet this need for analysis, inventories have been prepared from two points of view: one, psychiatric; the other, psychological. The psychiatric inventories were intended to distinguish traits of insane persons, while the others were intended for use among the whole population. There is, of course, no sharp line of division here.

TABLE 94

Questionnaire Items on Neurotic Behavior

1. Do you get stage fright?
2. Do you have difficulty in starting conversation with a stranger?
3. Do you worry too long over humiliating experiences?
4. Do you often feel lonesome, even when you are with other people?
5. Do you consider yourself a rather nervous person?
6. Are your feelings easily hurt?
7. Do you keep in the background on social occasions?
8. Do ideas run through your head so that you cannot sleep?
9. Are you frequently burdened by a sense of remorse?
10. Do you worry over possible misfortune?
11. Do your feelings alternate between happiness and sadness without apparent reason?
12. Are you troubled with shyness?
13. Do you daydream frequently?
14. Have you ever had spells of dizziness?
15. Do you get discouraged easily?
16. Do your interests change quickly?
17. Are you easily moved to tears?
18. Does it bother you to have people watch you at work, even when you do it well?
19. Can you stand criticism without feeling hurt?
20. Do you have difficulty making friends?
21. Are you troubled with the idea that people are watching you on the street?
22. Does your mind often wander badly so that you lose track of what you are doing?
23. Have you ever been depressed because of low marks in school?
24. Are you touchy on various subjects?
25. Are you often in a state of excitement?
26. Do you frequently feel grouchy?
27. Do you feel self-conscious when you recite in class?
28. Do you often feel just miserable?
29. Does some particular useless thought keep coming into your mind to bother you?
30. Do you hesitate to volunteer in a class recitation?
31. Are you frequently in low spirits?
32. Do you often experience periods of loneliness?
33. Do you often feel self-conscious in the presence of superiors?
34. Do you lack self-confidence?
35. Do you find it difficult to speak in public?
36. Do you often feel self-conscious because of your personal appearance?
37. If you see an accident, are you quick to take an active part in giving help?
38. Do you feel that you must do a thing over several times before you leave it?
39. Are you troubled with feelings of inferiority?
40. Do you often find that you cannot make up your mind until the time for action has passed?
41. Do you have ups and downs in mood without apparent cause?
42. Are you in general self-confident about your abilities?
43. (Above the median, A. C. E. Psychological Examination.)

(Mosier, 1937, p. 280. By permission of the Editor, *Psychometrika*.)

A. Psychiatric Classification

One of the most elaborate of the inventories designed to show psychiatric syndromes is the Humm-Wadsworth (1934) temperament scale. Following a widely used psychiatric classification, eight types are distinguished: Normal, Hysteroid, Anti-social, Manic-cycloid, Depressive-cycloid, Autistic-schizoid, Paranoid, and Epileptoid. The scale consists of 318 items to be checked as typical (yes) or not typical (no) of a person. Scores for each of these eight types of temperament are secured from keys. A check of this scale showed that nearly one half of the items failed to discriminate between the groups. The other half showed various discriminations. All items have been retained in this scale, however, because it was thought that the nondiscriminatory items probably influence answers on the rest. In addition to the analytical score, it is claimed that negativistic persons marked a much larger number of items *no,* and that suggestible persons marked *yes* much more frequently than the normal. A similar scale published by Cattell (1936) contains items for seven pathological conditions called: Neurasthenia, Anxiety Neuroses, Anxiety Hysteria, Conversion Hysteria, Compulsive, Epileptoid, and Paranoid. Both of these scales are intended for appraisal only with the most co-operative persons. They may also be used by clinicians for numerical evaluations of patients' traits. Both scales are valuable for furnishing a rapid systematic check of tendencies. They are not widely used as yet because of difficulties in interpreting various items and lack of standardization.

B. Psychological Theory

Another set of scales follows psychological theories of personality. These include scales of introversion-extroversion, ascendance-submission, sufficiency-dependence, and despondency-elation. Nearly 30 scales of this sort have been edited by various investigators. They are well illustrated by the work of Bernreuter (1931) and of Rundquist and Sletto (1936).

1. The Bernreuter Personality Inventory. This is doubtless the most widely applied test of its kind. It consists of 125 items, similar to those in Table 94. The items describe both adjustments and interests. Each item is to be checked *yes, no,* or *unable to answer with yes or no.* Four scores are obtained for each person from keys which were prepared on the basis of results from four previous tests: Thurstone's (1930) Personality Schedule of Neurotic Tendencies, Laird's (1925) Inventory of Extroversion-Introversion, Allport's (1928) Ascendance-Submission Scale, and Bernreuter's Test of Self-Sufficiency. These four tests and the Personality Inventory were administered to adults selected in part to represent extreme groups. Each item in the inventory was correlated with total scores on each of the four tests. The answers to each item were assigned points on the basis of these correlations, the higher the correlation, the more the points allotted. For instance, the answers to the item, "Do you day-dream frequently?" were given plus and minus values as follows:

	Test			
Answer	*Neurotic*	*Introversion*	*Dominance*	*Self-Sufficiency*
Yes	5	3	−1	1
No	−4	−4	1	−1
Doubtful	−2	0	2	−2

Total scale scores were secured by adding the figures in each column. These totals were found to correlate highly with the corresponding previous tests. Thus, Bernreuter's score for neurotic tendencies correlated .94 with Thurstone's schedule. Laird's and Bernreuter's introversion scores correlated .79. Allport's measure of ascendancy and Bernreuter's dominance correlated .81, and the two measures of self-sufficiency .89.

The Bernreuter scores show high split-half reliability, median .90. Their intercorrelations are interesting. Neurotic tendencies correlated .95 with introversion. This shows either that the same persons have both sorts of adjustments, or that the two scales are measuring approximately the same patterns of behavior. Neurotic tendencies correlated with ascendancy .81, and with self-sufficiency .35. The self-sufficiency scores showed a low correlation with the others since self-sufficiency cuts across both ascendancy and introversion.

These high intercorrelations led Flanagan (1935) to make a factorial analysis of Bernreuter's scores from 305 eleventh-grade boys. He used Hotelling's method of principal components. Two factors appeared to account for the intercorrelation of the four Bernreuter scores. The first, a large factor in the test, is a combination of neurotic, introversion, submission, and low self-sufficiency items. Flanagan named this *lack of self-confidence*. The second factor, a much smaller one, was called *sociability*. Flanagan constructed two new scoring keys to aid in appraising these modes of behavior.

The stability of responses on the Bernreuter scale was studied by Farnsworth (1938). Retests after one, two, and three years showed no significant shifts in individual centile ranks, and there were high retest correlations. For the average person, 71 per cent of single items were answered in identical fashion after an interval of one year, 65 per cent after two years, and 65 per cent after three years.

2. THE MINNESOTA SCALE FOR THE SURVEY OF OPINION. In constructing scales to be used for measuring the effects of the depression on personality and family life of young people, Rundquist and Sletto (1936) drew most of the preliminary items from their own logical considerations. Care was exercised to secure informal language and clarity of statement. The personal pronoun was avoided since they believed that impersonal statements of majority opinion

might be more frank. Statements of positive and negative sorts were made in equal numbers, the positive usually expressing optimistic or socially acceptable views. The items were submitted to groups of students for criticism. Statements containing the words *all, always, none,* and *never* were considered to be unsatisfactory, for students sometimes disagreed with this type of statement although agreeing with the main conclusion of the item.

The statistical selection of items included two considerations: reliability and discriminative ability. The reliability of each item was secured by giving the scales twice within a week. Items which showed wide fluctuations were discarded. The discrimination of an item was measured by noting differences in answers to it by the highest and lowest quarter of a group. The items finally selected usually showed the largest differences. More negative than positive statements were found to be discriminative.

The final inventory of 132 items contains scales for measuring six aspects of adjustment:

1. Morale: feelings of inability to cope with one's problems
2. Social inferiority: feelings of inability to succeed in association with others
3. Family: ideas about pleasantness and intimacy of family life
4. Law: attitudes toward legal institutions
5. Economics: conservatism and radicalism
6. Education: belief about the values of education

Each scale consists of 22 items, and each item is to be marked on a 1 to 5 scale, using the words *strongly agree, agree, undecided, disagree,* and *strongly disagree.* Table 95 prints the first 18 items, of which Nos. 1, 7, and 13 are scored for morale, Nos. 2, 8, and 14 for social inferiority, etc.

Norms are available for 1,000 young persons – 400 from college, 200 from regular high schools, and 400 from continuation high school classes. No significant differences between high school and college students were found. Individual profiles can be made to show one's position in this

TABLE 95

Minnesota Scale for the Survey of Opinions

Directions: READ EACH ITEM CAREFULLY AND UNDERLINE QUICKLY THE PHRASE WHICH BEST EXPRESSES YOUR FEELING ABOUT THE STATEMENT. Wherever possible, let your own personal experience determine your answer. Do not spend much time on any item. If in doubt, underline the phrase which seems most nearly to express your present feeling about the statement. WORK RAPIDLY. Be sure to answer every item.

1. THE FUTURE IS TOO UNCERTAIN FOR A PERSON TO PLAN ON MARRYING.
   Strongly agree ^{5} Agree ^{4} Undecided ^{3} Disagree ^{2} Strongly disagree ^{1}
2. AFTER BEING CAUGHT IN A MISTAKE, IT IS HARD TO DO GOOD WORK FOR A WHILE.
   Strongly agree ^{5} Agree ^{4} Undecided ^{3} Disagree ^{2} Strongly disagree ^{1}
3. HOME IS THE MOST PLEASANT PLACE IN THE WORLD.
   Strongly agree ^{1} Agree ^{2} Undecided ^{3} Disagree ^{4} Strongly disagree ^{5}
4. THE LAW PROTECTS PROPERTY RIGHTS AT THE EXPENSE OF HUMAN RIGHTS.
   Strongly agree ^{5} Agree ^{4} Undecided ^{3} Disagree ^{2} Strongly disagree ^{1}
5. THE GOVERNMENT SHOULD TAKE OVER ALL LARGE INDUSTRIES.
   Strongly agree ^{5} Agree ^{4} Undecided ^{3} Disagree ^{2} Strongly disagree ^{1}
6. A MAN CAN LEARN MORE BY WORKING FOUR YEARS THAN BY GOING TO HIGH SCHOOL.
   Strongly agree ^{5} Agree ^{4} Undecided ^{3} Disagree ^{2} Strongly disagree ^{1}
7. IT IS DIFFICULT TO THINK CLEARLY THESE DAYS.
   Strongly agree ^{5} Agree ^{4} Undecided ^{3} Disagree ^{2} Strongly disagree ^{1}
8. IT IS EASY TO EXPRESS ONE'S IDEAS.
   Strongly agree ^{1} Agree ^{2} Undecided ^{3} Disagree ^{4} Strongly disagree ^{5}
9. PARENTS EXPECT TOO MUCH FROM THEIR CHILDREN.
   Strongly agree ^{5} Agree ^{4} Undecided ^{3} Disagree ^{2} Strongly disagree ^{1}
10. A PERSON SHOULD OBEY ONLY THOSE LAWS WHICH SEEM REASONABLE.
    Strongly agree ^{5} Agree ^{4} Undecided ^{3} Disagree ^{2} Strongly disagree ^{1}
11. LABOR SHOULD HAVE MUCH MORE VOICE IN DECIDING GOVERNMENT POLICIES.
    Strongly agree ^{5} Agree ^{4} Undecided ^{3} Disagree ^{2} Strongly disagree ^{1}
12. THE MORE EDUCATION A MAN HAS THE BETTER HE IS ABLE TO ENJOY LIFE.
    Strongly agree ^{1} Agree ^{2} Undecided ^{3} Disagree ^{4} Strongly disagree ^{5}
13. THE FUTURE LOOKS VERY BLACK.
    Strongly agree ^{5} Agree ^{4} Undecided ^{3} Disagree ^{2} Strongly disagree ^{1}
14. IT IS DIFFICULT TO SAY THE RIGHT THING AT THE RIGHT TIME.
    Strongly agree ^{5} Agree ^{4} Undecided ^{3} Disagree ^{2} Strongly disagree ^{1}
15. ONE OUGHT TO DISCUSS IMPORTANT PLANS WITH MEMBERS OF HIS FAMILY.
    Strongly agree ^{1} Agree ^{2} Undecided ^{3} Disagree ^{4} Strongly disagree ^{5}
16. IT IS ALL RIGHT TO EVADE THE LAW IF YOU DO NOT ACTUALLY VIOLATE IT.
    Strongly agree ^{5} Agree ^{4} Undecided ^{3} Disagree ^{2} Strongly disagree ^{1}
17. LEGISLATURES ARE TOO READY TO PASS LAWS TO CURB BUSINESS FREEDOM.
    Strongly agree ^{1} Agree ^{2} Undecided ^{3} Disagree ^{4} Strongly disagree ^{5}
18. EDUCATION HELPS A PERSON TO USE HIS LEISURE TIME TO BETTER ADVANTAGE.
    Strongly agree ^{1} Agree ^{2} Undecided ^{3} Disagree ^{4} Strongly disagree ^{5}

(Rundquist and Sletto, 1936. By permission of the University of Minnesota Press.)

sample of persons. The scores on the separate scales had reliability coefficients of approximately .85. The intercorrelations between these six scales for a sample of 500 young adults were:

	2	3	4	5	6
1. Morale	.511	.420	.518	.235	.534
2. Inferiority		.248	.262	.205	.231
3. Family			.388	.081	.303
4. Law				.191	.396
5. Economic Conservatism					.087
6. Education					

Morale scores were found to be most highly correlated with the other scores, and economic conservatism the least highly correlated.

In a small group of high school students correlations between these adjustment scores and mental ability are low with two exceptions. Morale correlated with honor points .506 for boys but not for girls, and attitude toward education correlated −.40 with IQ for girls but not for boys. The correlations of IQ's and honor points were .47 for boys and .56 for girls. The authors' original work should be consulted for a clearer analysis of their results from applying this scale to many parent-child, age, and social groups.

3. General Personality Scales. The last few years have witnessed the publication of a large number of self-appraisal scales which include a wide variety of items – some factual, and some descriptive of attitudes, interests, and modes of adjustment. Such scales have frequent application at the high school level, where personal and vocational counseling is attempted. Nearly all the large school systems have published short check lists which are rudimentary personality scales that aid in a rough classification of pupils according to their problems. Elaborate self-rating scales are illustrated by the work of Purdom (1932), Bell (1934), and Link (1936).

The Diagnostic Information Blank of Purdom (1932) consists of 84 items of a completion type. Some of these require long answers to such questions as, "What do you think about smoking and drinking?", but most of the items are answered by *yes* or *no,* as, "Do you feel that you have enough freedom to do what you want to do from day to day?" The answers are classified into 18 groups according to a key made from the consensus of opinion of a fairly large number of counselors. The result is a condensed record sheet which shows the types and acuteness of the difficulties which the pupil has indicated by his answers.

The Bell Adjustment Inventory (1934) (1938) consists of 160 items divided equally among five fields of adjustments — namely, adjustment to home, health, other persons, emotional disturbances, and occupation. The items were selected from a preliminary set on the basis of their discrimination between the upper and lower 15 per cent of individuals when ranked for total adjustment scores. Each item is to be checked *yes, no,* or *?* . Separate scores are available for each field as well as the total. The odd-even correlations ranged from .81 to .91 for the separate sections, and to .94 for the total scores. The intercorrelations of the separate sections ranged from −.06 to .51, median .24. This shows a marked degree of independence among these self-ratings. Two forms of this inventory are now available, one for adults, the other for students in high school and college. Both forms are reported to give fairly good diagnoses of poorly adjusted persons.

The Personality Quotient Test of Link (1936) consists of 150 items of the sort shown in Table 94 and also short lists of games, studies, group activities, and hobbies which are to be checked. Separate forms have been prepared for boys and girls. Each item is scored according to a key which assigns points of credit for items which describe personality assets. The check list answers count very little in the total scores. Separate scores may be obtained for four traits:

1. Social Initiative: Included here are such items as going to school affairs, paying compliments to people, holding elective offices, selling, making friends easily.
2. Self-determination: This trait is one of doing things which are considered right, even though they may be unpleasant, such as practicing the piano when one would rather go to the movies.
3. Economic Self-determination: Items in this scale refer to earning money by chores, selling tickets, saving, trading, making friends, belonging to clubs, leadership of any sort, learning people's names.
4. Adjustment to the Opposite Sex. This includes dancing, parties, and talking with persons of the opposite sex.

A total score is the sum of credits in the four subtests. It may be converted as follows into a Personality Quotient (PQ) which is similar in scope to an IQ. The difference between a pupil's score and the mean of his group is divided by the standard deviation of the group. This yields a SD score. The SD score is made comparable to an IQ by multiplying it by 17 and adding the product to 100. 17 is the approximate standard deviation of Terman's IQ's, and 100 is the usual mean.

The interpretation of these scores is not easy owing to the fact that each scale as well as the PQ is dependent upon unknown combinations of somewhat independent items. Thus, the same item may be credited in two different scales, and two persons may obtain the same total scores when one is high in Economic Self-determination and the other low in this sphere.

The odd-even reliabilities of the subtests ranged from .73 to .88. The intercorrelation of the subtests ranges from .26 to .70. Social Initiative correlated .704 with Economic Self-determination and .66 with Adjustment to the Opposite Sex. Self-determination was more independent, as shown in correlations of .261, .283, and .126 with the other three subtests. Grade norms show small increases in raw scores from the lower to the upper grades. Among boys nearly zero correlations were found with intelligence test scores and school achievement, but among girls the total PQ correlated approximately .27 with IQ and Self-determination correlated .34 with IQ and .30 with high school class ranks. A correlation of .04 between PQ's of 54 pairs of siblings was reported by Thompson (1938) while the same children showed a correlation of .35 between IQ's.

18

MODES OF ADJUSTMENT
Conduct Measures, Observations, and Laboratory Methods

I. INTRODUCTION

The last chapter gave descriptions of measures of modes of adjustment by free association and questionnaire methods. This chapter describes three other methods used to appraise modes of behavior: conduct measures, records of observations, and laboratory methods. The next chapter discusses the effectiveness of the various methods, and the factorial analyses of some results.

II. CONDUCT MEASURES

Undoubtedly the most elaborate quantitative reports of adjustment behavior from the point of view of ethics are those of Hartshorne and May (1928, 1929, 1930), which describe a six-year project made for the Institute of Social and Religious Research. These workers and their assistants summarized existing methods of appraising deceit, self-control, suggestibility, moral knowledge, reputation, and integration. With great ingenuity they improved old tests, devised new ones, and applied them to many different groups of school children to discover typical performances in various environments. The work of these men has inspired many others to continued research and test refinement. Although their schedules included many types of appraisal, conduct measures were very prominent. Fairly complete outlines of their work are given below to show the thoroughness of their investigations and to allow an evaluation of the various methods of procedure.

Conduct measures were secured by placing persons in natural situations, and then recording the number of times a pupil used unethical conduct to attain a goal. The test situations discussed immediately following were used to appraise deception.

A. Measures of Deception, Hartshorne and May (1928)

1. Copying Technique. This appraisal was conducted by having the pupils sit together in pairs. Each pair was then given two forms of multiple-choice tests that looked alike. Both forms contained the same words, but the choices were arranged in different orders, so that if one pupil copied from the other this might be detected by comparing their answers. This method was found ineffective because of the difficulty of distinguishing chance from copied errors, and also because it did not furnish equal opportunities to all who might desire to copy.

2. Self-Scoring Technique. In order to make an appraisal of deception in self-scoring, ordinary mental or achievement tests were given, collected, and scored without marking the test sheets. They were then returned to the pupils, who were asked to score their own tests from keys which were provided. Deception can be measured by the amount of difference between scores secured in these two ways. Tests of this sort, unless they have a wide range of difficulty, tend to present different motives for cheating among those with different abilities. This method proved to be fairly effective, but rather expensive.

3. Improbable Achievement. In this situation tests were used twice: once under supervision, in order that the pupil's actual achievement could be determined; and once without observation, in order that he could raise his score by cheating. A wide variety of these tests were used, such as:

1. Puzzle solution in which the rules of the game may be broken
2. Weight discrimination in which seven pill boxes have the weights printed on the bottom
3. Peeking tests in which mazes or other figures are to be traced with the eyes closed, and in parlor games such as "pinning the tail on the donkey"
4. Two equivalent forms of mental tests in which the answer sheet is furnished with only one form
5. Two vocabulary tests, one taken home
6. Athletic tests (self-recorded measures of strength of grip, lung capacity, chinning, and broad jump are used, one test in public, one test in private)
7. Potato race (cheating is measured by the number of times a child broke the rules by picking up two potatoes)
8. Three trial-of-speed tests, such as cancellation of A's, digit symbol substitution, and making dots in small squares

4. Stealing. This group of tests included two types of appraisal:

1. Tests in which coins are used in puzzles or problems (The score indicates the number of coins not returned to the experimenter.)
2. Situations in which a storekeeper returns too much change. or in which a purse with money is found

5. Lying. In order to measure lying to escape disapproval, the pupils were asked to answer questions concerning their cheating activities on the tests just described. Lying to win approval was evaluated by a questionnaire on socially approved activities (see Table 96).

Deception was inferred by a comparison of the actual (or probable) and the self-reported achievements. If the self-report were considerably beyond the limits of probable achievement, the pupil was discredited.

TABLE 96

Lying to Win Approval

This method consists of a series of rather personal questions. There are many specific acts of conduct which on the whole have rather widespread social approval, but which at the same time are rarely done. The questions revolve around situations of this sort.

The test is in two forms. Each form contains 36 questions.

CEI ATTITUDES SA

Form One

Name ______________________ Date ______________________

School ______________________ Grade ______________________

Answer the following questions by underlining YES or NO. If your answer is YES, draw a line under YES. If your answer is NO, draw a line under NO. Please answer every question.

1.	Did you ever accept the credit or honor for anything when you knew the credit or honor belonged to someone else? .	YES	[✓ NO]	1.
2.	Did you ever act greedily by taking more than your share of anything?	YES	[✓ NO]	2.
3.	Did you ever blame another for something you had done when you knew all the time it was your fault?	YES	[✓ NO]	3.
4.	Do you usually report the number of a car you see speeding?	[✓ YES]	NO	4.
5.	Do you always preserve order when the teacher is out of the room?	[✓ YES]	NO	5.
6.	Do you report other pupils when you see cheating? . . .	[✓ YES]	NO	6.
7.	Did you ever pretend to understand a thing when you really did not understand it?	YES	[✓ NO]	7.
8.	Have you ever disobeyed any law of your country or rule of your school?	YES	[✓ NO]	8.
9.	Do you speak to all the people you are acquainted with, even the ones you do not like?	[✓ YES]	NO	9.
10.	Do you usually call the attention of people to the fact that you have on new shoes or a new suit or dress?	YES	[✓ NO]	10.

(And 26 more questions.)

(Hartshorne and May, 1928, p. 98. By permission of The Macmillan Co.)

B. Measures of Co-operation, Hartshorne and May (1929), Chapter III

1. SELF-OR-CLASS TEST. In this situation pupils were required to choose between working for the class or for them-

selves. Cash prizes in spelling were announced for both class and individual honors, but a child was not allowed to participate in both contests.

2. Allotment of Prize Money. Pupils were asked to vote whether the money should be distributed to members of the class, or given to the school or some hospital child.

3. Sharing of Equipment. Each pupil was given a box containing ten articles: a drinking cup, pencil sharpener, ruler, eraser, pen, penholder, double pencil, and three other pencils. He was then allowed to give whatever he wished to make up boxes for children who had no useful or pretty things.

4. Providing Material for Hospital Children. Each child was given a set of four envelopes and asked either to make or to find jokes, puzzles, pictures, and stories for sick children. He was also asked whether he would do this later, or whether he would like to help, but would not be able to do so.

5. Records of Service. Teachers made records from December to June on the actual activities of each pupil in the co-operation and service projects just described.

6. Verbal Description. This test was suggested by the ancient but vivid word sketches of the Greek writer Theophrastus. In order to get up-to-date materials, items representing co-operative or helpful acts were listed and arranged in rank order. The situations were then cast into short sketches of boy and girl behavior, ranked from most to least helpful, and given values from 9 to 0. Table 97 shows some of these sketches. Teachers were asked to match each pupil with the approximate portrait. A variety of this test is the "Guess Who" test of Maller (1932), which includes a wider variety of characteristics.

7. Check List of Traits. Two forms of 80 words, each describing co-operative acts, were applied to each child by at least two teachers. They also used five-point descriptive rating scales of co-operation and selfishness.

TABLE 97

Verbal Character Sketches

The final set, in order of helpfulness and with the score values based on the distribution of the pictures by the judges, is presented here:

9 — T. T is sincerely interested in promoting the happiness and welfare of everyone. His warm spirit of friendliness extends to all, no matter where they are or what their race, class, or creed may be. Even the slightest need stirs him to some friendly act. So strong is this purpose that he would endure the scorn of his fellows in order to help a struggling cause or to help persons in trouble, even though they might not seem to deserve help of any kind.

7 — F. F is quite sincere in wanting to be of help. To him it is not so much a matter of high purpose as of plain decency. When the cause seems to be a worthy one, he is very generous and would give a cherished possession or deny himself long-anticipated pleasures even for such seemingly remote interests as scholarship funds for older students in other lands or for hospitals for adults. He would persist in his helpful acts even though people thought he was very foolish to do so. He is prompt to offer his services to anyone obviously in trouble.

5 — N. N does not think much about being helpful as a general thing, but he has a kindly nature and his sympathies are easily aroused. When his emotions are appealed to, he is ready to help almost anyone or any cause, but would probably draw the line at helping people he disliked or who seemed foreign or undeserving. He always lends a hand when the occasion is obvious and, if the appeal is strong enough, would give away his own things or money he had saved for himself in order to help even such objects as a college for an underprivileged section of the country or a hospital for poor people. The disapproval of his parents or teachers would not dissuade him.

1 — L. L is not likely to help anyone at all unless he finds that others are doing so, especially his best friends. He'd chip in a little for flowers for a classmate or help his elder brother take care of his small sister if it did not take too much time from his play or if he didn't have to do too much work or spend more than a few cents. But for the most part he doesn't want to help, nor does he care whether he is regarded by others as helpful.

0 — P. Here is your thoroughly hard-boiled youngster. The person or cause needing help makes no difference at all to P; but if someone would call for it, he would let him take away some object he wanted to get rid of provided he would get some thanks or reward or prestige. Or he would leave something he didn't want to do anyway in order to engage in some helpful activity if it were very interesting on other grounds and if people would applaud him for doing so. But he does not care two cents about the object or about helpfulness as a duty or as the proper thing or as a way of winning favor or grace.

(Hartshorne and May, 1929, p. 82. By permission of The Macmillan Co.)

C. Measures of Persistence and Inhibition

1. STORY COMPLETION. In this test situation an exciting story was read to a class as far as the climax. The ending of the story was supplied in forms difficult to read, such as (*a*) with capital letters run together, (*b*) with small letters and capitals run together, and (*c*) with spaces between capital letters in the wrong places. As the child was instructed to draw vertical lines between words in order to facilitate reading, a score was readily determined by the number of words correctly marked off.

2. PUZZLES. Both mechanical and verbal puzzles were used. The scores were the time spent working.

3. LETTER COUNTING. Pupils were asked to count the letters in a page of pied type, and the time worked was recorded.

4. DISTRACTION TESTS. In these tests, lines of digits were to be added. The digits were printed among curious sets of pictures and lines. The results were compared with normal addition scores.

5. INHIBITION TESTS. These included situations where each child was given a small box of candy and asked not to touch it until after a series of tests. On another occasion a small combination safe was placed on each desk with instructions not to work at it until after a series of six tests. Scores on these tests were the number of times a child failed to comply by eating candy or manipulating the intriguing little safe before the set time.

6. RATINGS BY TEACHERS. Check lists of words describing self-control were used together with five-point descriptive scales of 8 items:

1. Control of temper
2. Control of attention
3. Control of laughter
4. Talking too much
5. Telling secrets
6. Impatience
7. Overindulgence in candy
8. Control of body movements

D. Tests of Moral Knowledge, Hartshorne and May (1930), Chapter IX

In order to discover what relation might exist between behavior and knowledge of right and wrong, a battery of information tests was collected. They tried to limit these tests to memory of subject matter, and to exclude capacity for ethical comparisons and generalizations. They included information on how people behave toward one another, constitutional rights, and causes and consequences of social behavior. The five types of tests illustrated below were included in this battery:

1. Cause and Effect Tests. This was a 30-item true-false test with such items as (p. 38):

Good marks are chiefly a matter of luck. T F
Success always comes from hard work. T F
God punishes bad people by making them sick. T F

2. Recognition. This was a multiple-choice form in which acts were to be classified as cheating (C), lying (L), stealing (S), other wrong acts (X), and not wrong (J) (p. 40).

Bullying younger children.
Using street car transfers that are out of date.
Apologizing for a misdeed when you are not very sorry.
Picking flowers in a public park.
Making up an excuse so as not to hurt someone's feelings.

3. Social-Ethical Vocabulary. This was a 150-item multiple-choice word knowledge test constructed by Gladys Schwesinger according to Thorndike's scaling technique. It is illustrated by (p. 43):

Bravery 1 folly 2 courage 3 livery
4 impertinence 5 humanity ________
Scoff 1 cold 2 angry 3 make fun of
4 extol 5 expound ________
Sumptuous 1 conceited 2 expensive 3 repast
4 meager 5 fairlylike ________

4. Free-Response Test. Presented with eight situations, the pupils were requested to write down everything that they could think of which might happen in each one. The score was the number of important and probable consequences written down (p. 43). Four of the situations were:

1. Whenever anyone picked on him he would go tell his teacher.
2. John accidentally broke a street lamp with a snow ball.
3. Ruth's folks had a crowded apartment, so they kept a lot of boxes of things on the fire escape.
4. Jim was anxious to make good marks at school, so he usually studied instead of going out to play with the other children.

5. Probability Test. This test contained the same eight situations listed in the last test and, in addition, printed lists of consequences which were to be checked (p. 45). One of these test items was:

Likely to Happen	Might, but Not Likely	Would Not Happen	John started across the street without looking both ways.
______	______	______	1. He got hit with an automobile.
______	______	______	2. He caused an accident to other people.
______	______	______	3. The traffic laws were changed so that boys could cross more safely.
______	______	______	4. It scared the auto drivers who saw him.
______	______	______	5. He was the cause of a driver being put in jail.
______	______	______	6. Two cars collided trying to avoid him.
______	______	______	7. He got across as safely as anyone.
______	______	______	8. He got confused when he looked up and saw a car coming.

In spite of their similarity, this test and the previous one showed a correlation of only .32.

E. Tests of Opinion, Hartshorne and May (1930)

In this group were included items which evaluated not only knowledge but also judgment on moral and social issues. A large variety of tests were included, samples of which are illustrated below.

1. DUTIES. In this test items were to be checked if they represented duties (p. 47):

1. To help a slow child with his lessons.
2. To read newspapers every day.
3. To call your teacher's attention to the fact that you received a higher grade than you deserved.
4. To keep a diary.
5. To sneeze when you feel like it.
6. To pray every day, etc.

Scores were established from class results and also from judgments of adults.

2. WHAT-YOU-WOULD-DO TEST. In this situation such multiple-choice items were presented as (p. 49):

1. If someone asks to borrow your pencil
   a. tell him it's broken.
   b. tell him you lost it.
   c. tell him you don't want to loan it.
   d. let him take it.
2. If someone steals your lunch
   a. steal another's lunch to even it up.
   b. report it to the teacher.
   c. cry about it.
   d. say nothing about it.

3. PROVOCATION TEST. In this situation each child was asked to rate the acts as right, wrong, or excusable (p. 50):

1. Helen noticed that nearly everyone in the class was cheating, so she cheated too.
2. Harry was a Christian boy. One day a Jewish boy called Harry a "dirty Christian." Harry knocked him down.

3. Charles did not want to play marbles for keeps, but the boys called him a "sissy," so he went ahead and played for keeps anyway.

4. Principles Test. This was a true-false test (p. 52):

1. To master oneself is a greater thing than to win a battle.
2. Clean speech is a sign of being a "goody-goody."
3. Obedience is of greater importance than honor.

5. Importance Test. In this case the items in the Probability Test listed above were ranked for importance.

F. Measures of Integration

The correlation of tests of honesty, co-operation, persistence, inhibition, moral knowledge, attitudes, and intellect tended to be so nearly zero that children in Grades V to VIII appeared to be loosely organized in these respects. It was possible, however, to calculate for each child the consistency of his scores. Figure 68 shows two profiles, one characterized by remarkable consistency, and the other by great variation. Both have the same average SD scores on 21 measures of deception. The index of integration chosen was the standard deviation of the SD scores. The boy in Figure 68 was found to have an index of .296, and the girl of 1.114. The reliability of such indices was found to be in the neighborhood of .411 for tests of deception. The correlation between indices and total honesty scores was .522, or .882 when corrected for attenuation.

Similar indices of integration were calculated on the basis of 23 measures of various sorts, including intellect, emotional stability, culture, honesty, co-operation, good citizenship, moral knowledge, opinion, teachers' records and ratings, and children's "Guess Who" tests. The average intercorrelation for these 23 measures was .30 for a group of 100 cases from 9 different classrooms. The sums of the 23 scores correlated .440 with indices of integration. Table 98 shows the correlations between integration indices and various sub-batteries. The conduct score appears to have less relation to the index

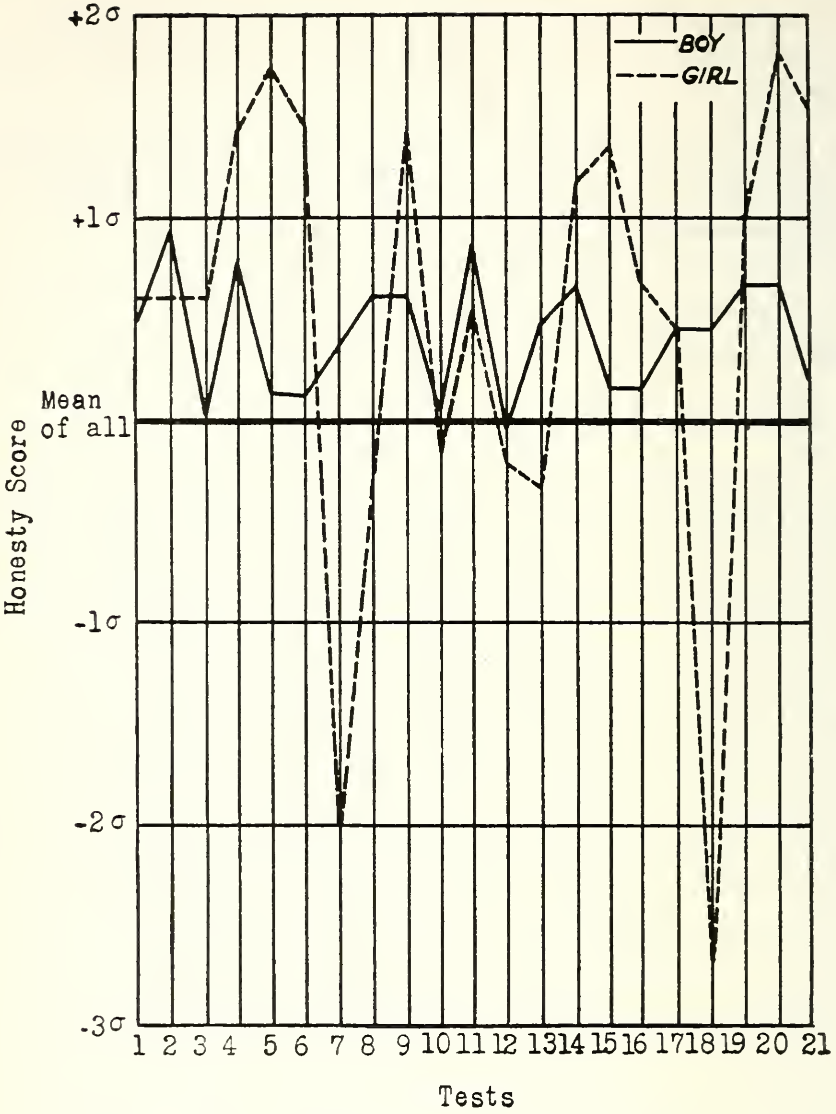


FIGURE 68. **Profile of Honesty Scores in Twenty-One Situations**

(Hartshorne and May, 1930, p. 290. By permission of the Macmillan Co.)

of integration than measures of knowledge, opinion, and reputation. Intellect as shown by the CAVI tests, cultural status, and emotional stability all show correlations below .20 with integrity, and age correlates nearly zero. There is a marked tendency for those who are more aggressive in de-

TABLE 98

Correlation of General Integration with Various Measures

Conduct			Knowledge		
	Raw r	*Corrected r*†		*Raw r*	*Corrected r*†
School honesty	.354	.61	Good citizenship	.504	.87
Service total	.179	.38	Information	.315	.53
Inhibition total *	.296	.74	Opinion A + B	.340	.58
Persistence total	−.066	−.13			
Reputation			**Ability and Status**		
	Raw r	*Corrected r*†		*Raw r*	*Corrected r*†
Teachers' marks	.361		CAVI (sigma)	.120	.20
Deportment	.232		Resistance to suggestion	.246	.49
Conduct record	.352		Emotional stability	.194	.37
"Guess Who"	.223		Self-Functioning	.289	.51
Total reputation	.400	.70	Age	−.041	−.06
			Sims (socio-economic)	.138	.24
			Burdick (culture)	.184	.33

* Omitting the Picture Inhibition Test. † Corrected for attenuation.

(Hartshorne and May, 1930, p. 351. By permission of the Macmillan Co.)

ception to have lower indices of integrity than do those who are more aggressive in service and persistence.

The interpretation of this material is difficult, since all one has to go on is a pupil's relative position in his class on 23 tests and ratings. It was shown that a pupil might have a low integration index and still be rated as well adjusted by a teacher. The index of integrity is thus not an indication of consistent individual behavior, but rather of consistency of measures of abilities and adjustment among a group of persons. The authors conclude that motives, interests, and attitudes are dependent upon a variety of environmental and natural forces, some of which tend to make persons alike and some different from one another. Opinions and behavior scores of parents and associates were found to correlate fairly well with scores of pupils.

III. RECORDS OF OBSERVATIONS

Records of direct observation of persons are among the most valuable indicators of adjustments. Three procedures are commonly used. In one a continuous record is kept over 24 hours a day of all important episodes and their stimuli. This sort of record can only be made in a hospital, or camp, or home situation where trained observers are available. The second procedure is called time-sampling. Many scattered periods, usually from 5 seconds to 5 minutes, are used for observing a particular person. The third procedure records observation in a single interview or test period. All three of these procedures will be described.

A. Continuous Records

One of the most thorough studies of a 24-hour record of a family situation was reported by Bühler (1939), who sent trained workers into private homes. The work of Newcomb (1929) is an outstanding example of analysis of detailed daily records made by counselors in a summer camp. In order to check the value of the general type concept of extroversion-introversion, he selected 30 items usually found in extroversion scales. Each item was printed with a four-point scale of specific behavior, as illustrated in Table 99. The specific behavior consistency of each boy on each situation was fairly high from day to day, mean r equaled .78, but considerable variation took place over longer periods.

From the same items he determined "trait" consistency by grouping together those specific behavior situations which seemed to involve similar activities or goals. Thus, under *volubility* were clustered instances of loud threats, loquacity, chattering, boasting, and announcing intention. An average intercorrelation of .26 was found among specific situations which were thus grouped together. This was not considered

TABLE 99

Daily Behavior Record

Numbers indicate how many times a day the behavior was observed.

Date ________________ Counselor ____________________

NAME OF BOY:	John	Henry	Pete	Ed	Frank
Did he show confidence in his own abilities?					
boasted loudly of greater abilities than he had	2				
spoke confidently of ability he really had	1		4	2	
expressed lack of confidence in own abilities		6		1	1
hesitated even to try his ability		3		1	
Did he take the initiative in organizing games?					
insisted on doing the organization himself	4				
gave constant advice to the leader	6	3			2
helped to plan, but loyal to the leader		1		1	
let others do all the planning — follower			4	4	
Did he submit to criticism or discipline from counselors?					
resisted violently, or fought back					1
retorted angrily, sarcastically, with threats	2				2
showed resentment by mumbling or sullenness	6	0	0		
accepted it quietly and in good spirit				0	

A new record sheet was used each day to avoid the possibility that the counselor might be influenced by previous records.

(After Newcomb, 1929, p. 21. By permission of the Bureau of Publications, Teachers College, Columbia University.)

to be evidence of the existence of a central factor. Similarly, 10 traits were distinguished and scored: energy output, ascendency toward authority, ascendency toward other boys, volubility, seeking limelight, interest in environment, impetuousness, social forwardness, ease of distraction, and preference for the group. The mean intercorrelation of total trait scores was only .37, a fact which led Newcomb to conclude that there was no evidence here for an extroversion-introversion type of person.

Ratings of behavior, based on several months of observation, have come to hold an important place in surveys of school populations. Two types of schedules are illustrated in the widely used form prepared by Haggerty, Olson, and Wickman (1930). Schedule A asked the rater to indicate the frequency of occurrence on a four-point scale of 15 types of behavior problems, such as cheating, lying, temper outbursts, speech difficulties, imaginative lying, sex offenses, and truancy. In total scores the more serious problems as shown from clinical records were given three times as much weight as the less serious. Schedule B has 35 items in four divisions. Items in the first division referred to intellectual aspects of behavior, in the second to physical, in the third to social, and in the fourth to emotional aspects. Each of these items, examples of which are shown in Table 100, is in the form of a five-point rating scale. The points were assigned by finding the average scores on Schedule A for pupils who received a particular rating on an item in Schedule B. Table 100 shows that the subdivisions of the item "How intelligent is he?" were assigned points on the basis of lack of intelligence. The reason for doing this is that the groups having the lowest intelligence rating also had the greatest number of behavior problems. In Item 2, "Is he abstracted or wide-awake?" the points do not correspond directly to amounts of this trait, for it was found that the most alert had more behavior difficulties than two less alert groups.

Schedule B was found to correlate .60 with total scores on Schedule A, and a composite score using both schedules correlated .76 with the frequency with which children were sent to the elementary-school principal by teachers or monitors. The self-consistency of teachers when rating twice within a short period was found to be .86 using Schedule B, and the split-half reliability of single rating was .92. Ratings of two elementary-school teachers commonly correlated .60 with each other when rating small groups of children.

An unusually thorough schedule designed to appraise

TABLE 100

Behavior Rating Scale

Directions for Using

SCHEDULE B

1. Do not consult anyone in making your judgments.
2. In rating a person on a particular trait, disregard every other trait but that one. Many ratings are rendered valueless because the rater allows himself to be influenced by a general favorable or unfavorable impression that he has formed of the person.
3. When you have satisfied yourself as to the standing of this person in the trait on which you are rating him, indicate your rating by placing a cross (X) immediately above the most appropriate descriptive phrase.
4. If you are rating a child, try to make your ratings by comparing him with children of his own age.
5. The masculine pronoun (he) has been used throughout for convenience. It applies whether the person whom you are rating is male or female.
6. In making your ratings, disregard the numbers which appear below the descriptive phrases. They are for use in scoring.

DIVISION I

1. How intelligent is he?

Feeble-minded	Dull	Equal of average child on street	Bright	Brilliant
(5)	(4)	(3)	(2)	(1)

2. Is he abstracted or wide-awake?

Continually absorbed in himself	Frequently becomes abstracted	Usually present-minded	Wide-awake	Keenly alive and alert
(5)	(4)	(2)	(1)	(3)

3. Is his attention sustained?

Distracted: jumps rapidly from one thing to another	Difficult to keep at task until completed	Attends adequately	Is absorbed in what he does	Able to hold attention for long periods
(5)	(4)	(3)	(1)	(2)

DIVISION II

8. Is he slovenly or neat in personal appearance?

Unkempt, very slovenly	Rather negligent	Inconspicuous	Is concerned about dress	Fastidious, foppish
(5)	(4)	(2)	(1)	(3)

9. How does he impress people with his physique and bearing?

Repulsive	Makes an unfavorable impression	Generally unnoticed physique and bearing	Makes a favorable impression	Excites admiration
(5)	(4)	(3)	(2)	(1)

(Haggerty, Olson, and Wickman, 1930. By permission of the World Book Co.)

causes as well as symptoms is the Detroit Scale for the Diagnosis of Behavior Problems by Baker and Traphagen (1935). It requires ratings by a trained investigator on 66 items. Each item is rated on a five-point scale, 1 *very poor,* 2 *poor,* 3 *fair or average,* 4 *good,* and 5 *very good.* The ratings are to be based on direct observations, medical and school records, and questioning of both parents and children. The ratings are described in detail for each item. Thus for Item 25, Later Recreational Facilities, the pupil is asked, "How do you amuse yourself after school, or during vacation? What do you have to play with?" and the parents are asked, "What does he have to play with and how does he spend his time?" Six additional questions for parents are also suggested (p. 65):

> What specific things does he have to play with? With what does he amuse himself?
>
> Is there a purpose in his play or is it spasmodic and of no educational value?
>
> Does he play at having shows or at make-believe school? Does he play ball, skate, or build model houses?
>
> Does he have a pet and does he care for it well?
>
> Does he get any companionship with his parents through sharing play and recreation with them?
>
> Does he complain that time hangs heavy on his hands?

Finally, for rating this factor, the following descriptive scale is given (p. 66):

POINTS	FACTORS
5	Has a few well-organized things to do and desirable companions in doing them.
4	Has good attitude but little opportunity to play or express himself constructively with companions.
3	Except for one or two things or for short periods, does not know what to do.
2	Has little to play with; is alone most of time; has no particular purpose or drive to activities.
1	No purpose in play activities; no place to play; playthings in poor shape; plays mostly away from home; no interest on parents' part.

The 66 items are classified under five headings: (1) health and physical factors, (2) personal habits and recreational factors, (3) personality and social factors, (4) parental and physical factors of the home, and (5) home atmosphere and school factors. A summary sheet shows the rating assigned to each item, the number of items which were given each rating, and also a total score obtained by simply adding all the credits together. The total score is then transmuted into a letter grade according to a table, which gives the letters nearly the same significance as that found in the usual class in elementary school or in the United States Army Alpha tests. The following summary of a case (from pp. 336–338) will illustrate the procedure:

CASE 5 – Behavior Rating C–

D. M., considered as a behavior case, was a fifteen-year old boy, with a score of 218. The distribution of his items is as follows:

SUMMARY OF FACTORS

Category	*Number*	*Weighted Score*
Very Poor	5	5
Poor	13	26
Fair	22	66
Good	9	36
Very Good	17	85
	66	218

He was considered as an extreme behavior case in his locality, which was somewhat above the average. The five very poor items were as follows: Early health, Item 1, seems to have had a considerable influence on his problem, as his father particularly was very indulgent to him on account of his health. Item 29, on anger and rage, shows that he was inclined to yield to fits of anger easily and was frequently picked on by other children. Item 36, vocational interests, seems to be very negative on account of his health, poor scholastic record, and too much parental indulgence. Item 48 shows that the father is an unskilled laborer, doing truck driving and janitorial jobs, with some possible feelings of inferiority about it. As to Item 65, D. M.'s record was consistently poor, with D's and E's, which reflect in part his poor vocational interests.

The items rated poor are as follows: Degree of vision defect, Item 7 – he always had very poor vision in one eye, but the other eye was approximately normal. His father was rather opposed to making any adjustments to improve his vision. Item 5, accidents – he was hit by an automobile, with minor injuries, and two years before he had broken his nose in a bicycle accident. Item 15 – he was very unattractive in appearance because of his crossed eyes. Items 16 and 17 – his early care and present care were poor. Items 19 and 20 – he ate meals at irregular times, hurried his eating, and was rather fussy and disagreeable about what he ate. Item 33 showed an IQ of 76 on the Stanford-Binet Test, which explains some of his difficulty in making adjustments to regular grade standards. Item 35, initiative and ambition, showed him not very ambitious. Items 40 and 41, education of both parents, showed that they had completed only two or three grades and in European countries. Item 61 – his discipline was inconsistent and modified by pity on account of his physical health. Item 66, his attitude toward the school – in this respect he was poor because the regular school work was too difficult for him. The parents refused to place him in any type of special class.

It seems that in D. M.'s case, in the next two or three years of school, vocational and social adjustments would be critical factors in determining whether he would make a successful final adjustment or whether this combination of factors would tend to carry him downhill.

Baker and Traphagen reported correlations between total scores and each item on two groups: 189 behavior cases, 180 non-behavior cases. The total scores on the first group correlated most highly with Family Recreation, .699; Ideals of the Home, .657; Conditions of Eating, .562; Time of Sleeping, .555; Father's Age at Birth of Child, .551; Economic Status, .526; Child's Intelligence, .525; Mother's Personality, .495; and similar items. Among the non-behavior cases, the total scores correlated most highly with Scholarship, .663; Later Recreational Facilities, .616; Personal Hygiene, .583; Initiative and Ambition, .569; Discipline, .558; Mother's Intelligence, .542; Conditions of Eating, .526; Father's Intelligence, .512; and Mother's Personality, .500; and Early Self-Care, .507.

The following items had low correlations with total scores among the behavior problem children: Children's Diseases, .03; Size for Age, .11; Speech Defects, .13; Infectious Diseases, .182; Early Self-Care, .175; Mother's Health, .200; and Broken Home, .254. Among the well-adjusted group the total scores showed low correlations with General Home Atmosphere, .117; Children's Diseases, .05; Defects of Vision, .154; Size for Age, .230; Economic Status, .220; and Child's Intelligence, .275.

These figures illustrate very well the results of correlation analyses of different groups of pupils. The differences among poorly adjusted pupils seemed to be largely dependent upon the interplay of living conditions and the personality of parents and associates. Among the well-adjusted group, total scores corresponded to differences in scholarship, recreations, and personal hygiene. When both groups of pupils were combined, a correlation of .91 was found between total scores and a combination of the five items: Discipline, Child's Attitude toward the Home, Parents' Attitude toward Child, General Behavior, and Scholarship.

B. Time-Sampling

Time-sampling methods have been found to be particularly well adapted to the study of modes of adjustment. A large number of studies in this field have been reviewed by Olson and Cunningham (1934), who find the method usually reliable. A good illustration is the work of Goodenough (1930), who reported 25 one-minute observations on 33 nursery-school children. The scores for each child were the amounts of time spent in each activity. The odd-even reliabilities of observations of laughter, sociability, leadership, and physical activity were approximately .80. Reliabilities of observation of compliance and talkativeness were approximately .50 and .60. The correlations between sociability, leadership, and talkativeness were generally above .60. Leadership also showed small positive relationships with height, weight, chronological age, and mental age, but zero relationships with sex, physiological tests, and size or status of family.

C. Records of Interviews and Test Situations

Employment officers have long relied upon their judgments of an applicant's modes of adjustment during an interview, often making an important decision with very little evidence. Doubtless, capable interviewers can size up a person's moods and mannerisms during an interview fairly well, but the brevity of the interview and the bias of the interviewer may lead to fragmentary or erroneous records. In order to help interviewers and oral examiners make accurate and systematic appraisals, a rating form shown in Table 101 was prepared. This form is typical of many that are used in ratings of employees who have been in service for some time. It consists of 9 general aspects, each of which is described by a brief paragraph and followed by a graphic and descriptive rating scale. This particular form, which has been made so that it can be scored by a machine, gives a profile for each individual.

Elaborate reports of adjustments during test situations have been standardized and codified by a number of authors. Table 102 shows a rating scale by Vernon which has been found to give fairly consistent results when two judges rated the same interviewee.

IV. LABORATORY METHODS

A laboratory approach to the measurement of persistence was reported by Howell (1933), who tried to get at an element of personality which seemed independent of strength or skill. He used tasks which involved constantly increasing distress from one of several sources, such as fatigue from gripping a hand dynamometer, forcing a needle or blunt peg into the flesh, heat from an electric grill, and electric shock. The procedure was to apply the stimulus gradually until the student gave the signal to stop, indicating that he had reached his limit of endurance. Individual tests were made with 102 subjects who were told that they were competing for rank and that their names and ranks would be posted. Twenty-four of these students had been previously threatened with dismissal from college for academic failures.

The results of the tests involving strength and pressure were found to correlate .44 with one's weight; hence a corrected score on these tests was made to equalize weight. The reliability of the battery estimated by correlating the odd-even tests ranged from .19 to .85 for small samples. The intercorrelations of the test ranged from .18 to .72, median approximately .47. The willingness to endure the needle and the heat had low correlation with each other and with the painful pressure situations. The total scores on these persistence tests were found to correlate .81 with the Volitional-Perseverance Section of the Downey Will-Temperament Profile, and .44 with the Allport (1928) Ascendance-Submission Test. Persistence scores correlated .18 with being a male, when the influence of weight was partialled out. The correlation between Ohio State University Intelligence Test scores and persistence scores was .10, but the latter correlated with grades in college .44. The prediction of grades from intelligence tests was .51. This was raised to .64 when persistence scores were combined with intelligence, by a multiple correlation.

Another interesting laboratory technique, developed by Luria (1932), has been applied widely in Russia and more recently in America. He made two records of hand movements in conjunction with free association responses to words. The dominant hand was required to press a bulb to which a pressure recorder was attached. The other hand, about which nothing was said, was allowed to rest on a

TABLE 101

Rating Form for Use of Interviewers and Oral Examiners
– 2 – 1938

INSTRUCTIONS: Ask yourself how this applicant compares with those who are doing work of this kind. Consider whether his voice, appearance, etc., would be a liability or an asset in such a position. Rate him by making a check (✓) at that point on each scale where, in your judgment, the applicant stands. Rate the following traits:

1. **VOICE AND SPEECH.** Is the applicant's voice irritating, or pleasant? Can you easily hear what he says? Does he mumble, or talk with an accent which offends or baffles the listener? Or is his speech clear and distinct, his voice so rich, resonant and well-modulated that it would be a valuable asset in this position?

2. **APPEARANCE.** What sort of first impression does he make? Does he look like a well-set-up, healthy, energetic person? Has he bodily or facial characteristics which might seriously hamper him? Is he well-groomed or slovenly? Erect or slouchy? Attractive or unattractive in appearance?

3. **ALERTNESS.** How readily does he grasp the meaning of a question? Is he slow to apprehend even the more obvious points, or does he understand quickly, even though the idea is new, involved or difficult?

4. **ABILITY TO PRESENT IDEAS.** Does he speak logically and convincingly? Or does he tend to be vague, confused or illogical?

5. **JUDGMENT.** Does he impress you as a person whose judgment would be dependable even under stress? Or is he hasty, erratic, biased, swayed by his feelings?

6. **EMOTIONAL STABILITY.** How well poised is he emotionally? Is he touchy, sensitive to criticism, easily upset? Is he irritated or impatient when things go wrong? Or does he keep an even keel?

7. **SELF-CONFIDENCE.** Does he seem to be uncertain of himself, hesitant, lacking in assurance, easily bluffed? Or is he wholesomely self-confident and assured?

8. **FRIENDLINESS.** Is he a likeable person? Will his fellow-workers and subordinates be drawn to him, or kept at a distance? Does he command personal loyalty and devotion?

9. **PERSONAL FITNESS FOR THE POSITION.** In the light of all the evidence regarding this person's characteristics (whether mentioned above or not) how do you rate his personal suitability for work such as he is considering? Recalling that it is not in his best interest to recommend him for such a position if he is better suited for something else, would you urge him to undertake this work? Do you endorse his application?

Fuller instructions and space for comments on applicant's behavior will be found on the back of this sheet.

TABLE 101 *(Continued)*

Applicant's Name or
Identification Number ....Date..............................

Kind of work for which his
suitability is appraised ....

Irritating or Indistinct	Understandable but rather unpleasant	Neither conspicuously pleasant nor unpleasant	Definitely pleasant and distinct	Exceptionally clear and pleasing
Unprepossessing or Unsuitable	Creates rather unfavorable impression	Suitable Acceptable	Creates distinctly favorable impression	Impressive Commands admiration
Slow in grasping the obvious. Often misunderstands meaning of questions	Slow to understand subtle points. Requires explanation	Nearly always grasps intent of interviewer's questions	Rather quick in grasping questions and new ideas	Exceptionally keen and quick to understand
Confused and illogical	Tends to scatter or to become involved	Usually gets his ideas across well	Shows superior ability to express himself	Unusually logical clear and convincing
Notably lacking in balance and restraint	Shows some tendency to react impulsively and without restraint	Acts judiciously in ordinary circumstances Might be hasty in emergencies	Gives reassuring evidences of habit of considered judgment	Inspires unusual confidence in probable soundness of judgment
Over-sensitive Easily disconcerted	Occasionally impatient or irritated	Well poised most of the time	Superior self-command	Shows exceptional poise, calmness and good humor under stress
Timid. Hesitant Easily influenced	Appears to be over-self-conscious	Moderately confident of himself	Wholesomely self-confident	Shows superb self-assurance
Keeps people at a distance	Does not easily attract friends	Approachable Likeable	Draws many friends to him	An inspirer of personal devotion and loyalty
Unsuited for this work. Not endorsed	Might do well. Endorsed with hesitance	**Endorsed**	Endorsed with confidence	Endorsed with enthusiasm

SIGNATURE OF RATER

This rating form prepared from suggestions furnished by W. V. Bingham.

Copyright, 1938, by the International Business Machines Corporation. All rights reserved.
Printed in U. S. A.
IBM FORM I.T.S. 11-00 A 195

TABLE 102

General Rating Scale

Name.........................Date................Examiner................

Activity

Excited, restless, unable to keep still
Quick and vivacious
Calm and deliberate
Inert and listless

Impulsive
Stable
Cautious
Inhibited

Poses, motor attitudes...

Tics.................Nail-biting.................Twitchings................

Fiddling with material.............clothes.............hands.............feet............

Peculiar expressions.......................... excessive wrinklings.........................

Movement

Fluent and graceful
Accurate and well controlled
Angular and awkward
Clumsy

Quick stride and movements
Slow stride and movements

Physique and Bearing

Impressive in bearing
Satisfactory impression
Unimpressive
Forceful, efficient, energetic, upright posture and gait
Slouching gait
Weak, inefficient movements and bearing
Plump (pyknic) proportions
Well and symmetrically proportioned
Thin (asthenic)

Healthy looking, well developed, and nourished
Unhealthy, feeble physique
Florid
Pale

Personal Appearance and Expression

Attractive and good-looking (positive reaction)
Pleasant
Uninteresting, indifferent attractiveness
Ugly and repulsive (negative reaction)
Strong expressiveness of face and gestures
Expressionless
Quick and strong sense of humour
Slow but sure
Unable to see humour
Mature, serious, philosophical
Immature, childish

Sensual..................
Effeminate...............
Frank
Secretive
Cheerful, optimistic
Depressed, melancholy
Excitable, irritable
Even-tempered
Calm, phlegmatic

Special Characteristics

....

....

....

Speech

Voice resonant, pleasing, well modulated
Hard, harsh, pinched
Expresses meaning directly, grammatically, with facility
Unable to express himself, ungrammatical
Garrulous, overtalkative
Rather voluble
Seldom speaks of own accord
Reticent, taciturn

Clear, fluent, distinct
Stutters, stammers
Accent...................
Brilliant in talking, wide vocabulary
Dull and stolid, narrow vocabulary

Personal Care

Fastidious in dress, overmanicured
Good taste, neat and clean
Passable and inconspicuous
Careless in dress and cleanliness
Slovenly and unkempt

TABLE 102—*Continued*

General Rating Scale

SELF-ASSERTION	
Pompous and overbearing	Decisive
Complacent	
Self-confident and possessed	Wavering
Self-critical and deprecatory	
Embarrassed, bashful, self-conscious	Contrasuggestible
Anxious, apprehensive	
Submissive, retiring	Suggestible
CO-OPERATIVENESS	
Willing to co-operate in every respect; enters into spirit	
Reserved and formal	
Constrained and suspicious, outside the situation	
Surly and hostile	
Scrupulous, punctual and regular in attendance, application	
Industrious	
Easy-going, indifferent	
Lazy and irregular	
ALERTNESS AND CONCENTRATION	
Intelligently attentive, wide-awake	
Concentrated	
Absent-minded	
Easily distracted, inattentive	
TEST REACTIONS: PLANNING	
Analytical	
Serious but unsystematic	
Trial and error	Profits by past experience
Haphazard	Repeats same mistakes
EMOTION	
Wild and unrestrained emotional behavior and remarks	
Wilful and childish reactions, capricious	
Some loss of self-control, and overt emotion	
Humorous and unconcerned	
Serious, philosophical	
Repressed and inhibited	

(Vernon, 1938, p. 56. By permission of the *British Journal of Educational Psychology*.)

movable plate to which a movement recorder was attached. For each verbal response, his results showed the intensity of two sorts of motor responses, voluntary and involuntary. From his results, measures on a scale of integration were available. At one extreme of the scale was complete coordination of speech and motor responses, and at the other extreme involuntary motor responses of considerable violence often preceded or followed the voluntary hand and speech reactions. Luria believed that the degree of integration depends upon the cortical mechanism. In the best integrated persons all conflicts were settled in the cortical centers before motor responses were made. The opposite was true in the most disorganized persons, where motor responses in many patterns were typical of conflict situations. The difficulty of the task could be varied experimentally. Marked tendencies toward disorganization were found in small children, feeble-minded persons, and hysterical patients. These same tendencies have been noted by Greene (1937a) in the results of aiming and comparison tests. He found that errors and disorganized behavior yielded ratios of integration which were fairly consistent throughout a variety of motor and mental tests.

Huston, Shakow, and Erickson (1924) repeated one of Luria's experiments with hypnotized subjects. They found that under hypnosis the subjects tended to exhibit few motor disturbances when they gave verbal responses related to the conflict. When normally awake, the same subjects gave fewer verbal responses, but more motor reactions. Consequently, the investigators suggested the hypothesis that there are various levels of discharge of nervous impulses. If the excitation created by a stimulus is not discharged verbally, it tends to spread to voluntary motor responses, and if not discharged there, to involuntary motor responses.

Olson and Jones (1931) and Sharp (1937) applied the Luria technique to find out what sorts of stimulus words cause the greatest responses in various types of subjects. Stimulus word lists containing 15 words each were prepared by Sharp to sample responses to five regions of activity: family, social, religion, health, and intellect. These words were distributed among 125 neutral words, and administered to college and high-school groups, scholastic failures, and stutterers. The following directions were used (p. 114):

> Rest your arms comfortably on the table. Place the first and second fingers of each hand in (the rubber which is stretched across the) cups in front of you. Now, I shall read a list of

words to you. As soon as you hear my words respond at once with the first word that comes to your mind. At the same time press your right hand. Your left hand is to remain quietly in the same position throughout the experiment.

Each response was scored to show the reaction time and the amount of voluntary and involuntary motor disturbance as shown by the right and the left hands, respectively. The test was repeated on a later day and the correlations were found to be fairly high between corresponding scores, as recorded in Table 103. This table indicates a greater con-

TABLE 103

Correlations between 1st and 2nd Trials of an Association-Motor Test

Test	*Reaction Time*	*Somatic Disturbances*
Total	.94 ± .01	.87 ± .02
Family	.95 ± .01	.84 ± .03
Social	.96 ± .01	.87 ± .02
Religious	.93 ± .01	.88 ± .02
Health	.90 ± .02	.67 ± .05
Intellect	.92 ± .01	.85 ± .03
Neutral	.75 ± .04	.68 ± .05

(Sharp, 1938, p. 118. By permission of the University of Iowa, *Studies in Child Welfare.*)

TABLE 104

Group Means of Reaction Time and Somatic Disturbances

Category	*Reaction Time (Sec.)*			*Somatic Disturbances*		
	GROUP: 1	2	3	GROUP: 1	2	3
Family	1.79	1.99	2.10	2.56	1.69	2.73
Social	2.14	2.14	2.55	6.58	4.17	4.23
Religious	1.60	2.25	2.43	2.78	3.25	2.73
Health	1.54	2.07	2.39	1.78	2.42	3.97
Intellect	1.88	2.12	3.00	3.62	2.90	13.73
Neutral	1.57	1.95	2.05	1.06	1.55	1.00

NOTE: Group 1. College Freshman Women
Group 2. High School Girls
Group 3. College Women Scholastic Failures and Stutterers

(After Sharp, 1938, p. 121. By permission of the University of Iowa, *Studies in Child Welfare.*)

sistency between trials for the verbal reaction times than for the motor disturbances, and also for the supposedly affectively toned words than for the neutral words. All reliabilities are so high that chance variations are fairly well ruled out.

The differences among group averages are also very interesting as shown in Table 104. Here it appears that average freshman girls in the university had shorter reaction times than either the scholastic failures or the high school girls. Words related to social situations caused the longest reaction times and also the greatest somatic disturbances among the average college freshman, whereas words related to intellectual situations produced these effects among the scholastic failures. The high school girls showed more verbal and nonverbal blocking from words concerned with religion and social activities. Individual profiles indicate quite clearly the existence of particular regions of stress which were in some cases not previously diagnosed.

The use of laboratory methods for the detection of lying has become fairly common in police groups. Early work by Larson and Marsden has grown today into fairly well-established practices which occasionally result in confessions, thus avoiding an expensive court trial. The usual procedure secures records of physiological responses during a period of interrogation. If peculiar responses are noted, deception is suspected. The responses most often measured are breathing, pulse, blood pressure, and electrical skin resistance, sometimes called the psychogalvanic reflex. Changes in these responses are recorded mechanically after each stimulus question or situation. A good appraisal of this technique is given by Larson and Haney (1932), Ruckmick (1938), and Woodworth (1938).

19

MODES OF ADJUSTMENT
Evaluation of Methods and Results

This chapter presents a comparison of methods, correlational analysis, and evaluation of items in studies of modes of adjustment.

I. COMPARISON OF METHODS

Because the methods which have been described in Chapters 17 and 18 seem to be designed to evaluate somewhat the same modes of adjustment, it is reasonable to inquire concerning their relative advantages. Three aspects will be discussed: economy, reliability, and validity.

A. Economy

The questionnaire and the self-rating procedures are by far the easiest to apply, for they require less time for administration than the others and are as easy to score. Ratings by others are a close second. Tests of conduct in the classroom and time-sampling of conduct elsewhere are more difficult to apply and score. Laboratory techniques require expensive apparatus and specialized training, whereas the other procedures are relatively less costly and easier to employ. If the same results could be secured by all methods, the self-ratings would doubtless be the most economical.

B. Reliability

The reliabilities reported in various ways are well illustrated by Table 105. This shows that two similar forms of measures of deception gave high correlations in all the tests of cheating by scoring one's own paper. Peeking when the eyes should be shut, faking a solution to a puzzle, and getting help from a dictionary were more variable activities.

TABLE 105

Reliabilities of Techniques Used for Measuring Deception

Types of Conduct	1*	2†
1. Copying from a key or answer sheet (3 tests)	.871	.863
2. Copying from a key or answer sheet (duplicating technique)(7 tests)	.825	
3. Adding on more scores (6 speed tests)	.825	.887
4. Peeping when eyes should be shut (3 co-ordination tests)	.721	.750
5. Faking a solution to a puzzle (3 tests)	.750	
(2 tests: Pegs and Fifteen Puzzle)	.620	
6. Faking a score in a physical ability test (4 tests)	.772	
7. Lying to win approval	.836	
8. Getting help from a dictionary or from some person on one test done at home		.240

* The correlations in this column are based on intercorrelations between similar forms, and predicted by the Spearman-Brown formula.

† The correlations in this column are based on retests.

(Hartshorne and May, 1928, p. 136. By permission of the Macmillan Co.)

Hartshorne and May (1930) reported that separate measures of *service* showed retest reliabilities near .90, but that the combined measures showed odd-even correlations between 6 subbatteries of only .78. The Moral Knowledge Tests showed a high correlation between two vocabulary forms, .94, but the Opinion tests showed less consistency due in part to smaller numbers of items and shorter times. When items and time were held constant all of these tests had high and similar reliability. Opinion Battery A showed considerably less consistency than Battery B, probably owing to the length of the tests.

These results emphasize the fact that consistent behavior is more evident in large samples of items than in small, because of the elimination of errors of measurement and of individual variations. Under the best conditions, self-ratings, conduct measures, and ratings by teachers of other pupils are nearly as reliable as are measures of skills, and all have high reliabilities when enough items are combined.

C. Validity

Definite answers can be given to questions of validity only when some true criterion of truth is available. Since there are no accepted criteria for most of the traits under consideration, one can only compare various appraisals to see how well they agree. A few comparisons are given below.

Maslow (1937) found from extensive studies of dominance that feelings of dominance as shown by self-ratings and observed dominant behavior were not closely related. Jarvie and Johns (1938) compared ratings on Bernreuter traits for students intimately known by the raters with the actual Bernreuter scores. From the very low correlations reported, they concluded that the scores had little value in the solution of adjustment difficulties.

A comparison by three questionnaires of normal adults and mental hospital cases who had been matched for age, schooling, occupations, and mental ability was made by Landis *et al.* (1935). The results showed that all three questionnaires, the Bernreuter Personality Inventory, the Page Questionnaire of Schizophrenic Traits, and the Woodworth Psychoneurotic Inventory, were not valuable for distinguishing normal from insane persons, although the various types of psychosis showed small differences in scores.

Burnham and Crawford (1935) marked the items on the Bernreuter test ten times according to a pair of dice. They found that chance marks secured in this fashion yielded fairly high neurotic and introverted scores.

Ratings on the Haggerty-Olson-Wickman Scale by teachers were compared with the results of simply asking teachers to name the boys and girls causing the most trouble in school. Of the lowest 10 per cent of the most poorly adjusted pupils as shown by H–O–W-scores, only about half were named by the teachers.

In a boys' camp Newstetter (1937) made a check on the validity of a preferred-associate ballot by actual observation of free time spent with various boys. He found a mean correlation of .73 between the two measures, and also reported that the best liked boys were not necessarily those who were cordial, but those who had some skill.

A valuable comparison of three types of appraisal was made by Hanks (1936). He secured six judges who read 50 autobiographies of about 1000 words which had been written as part of requirements of a college freshman English class. Each judge wrote a short analysis of each case showing the main types of responses, and also estimated the scores which had been made on three tests: (1) the Wilke Attitude Scale, (2) the Clark revision of Thurstone's Personality Schedule, and (3) the Deutsch Conformity Tests. The correlations between actual scores and estimates were all low, but slightly higher for the second test than for the other two. In most cases the SD of estimated scores was smaller than the SD of actual scores. In a further investigation Hanks instructed six judges to rate students after reading their autobiographies, using 28 items which concerned family relationships, adjustments in social, religious and economic spheres, and interests. These ratings were compared to self-ratings on the same items. The results showed an average of only 40 per cent correspondence which was raised to 46 per cent when the same judges were instructed in Adler's classification of types according to variations in co-operation and activity. This instruction did not lead to the use of Adler's types by all judges, but it did lead to a more logical and thorough approach to the task.

Hartshorne and May (1930) found that classmates' judgments, shown by a "Guess Who" test of reputation for desirable traits correlated .48 with teachers' judgments as represented by a weighted sum of grades in deportment, a conduct record, and a check list of traits. The teachers' check lists showed a correlation of .84 with the pupils' conduct records, and .49 with their deportment grades.

An inspection of these reports does not result in definite conclusions about the value of the several methods. We find much evidence that two methods designed to measure the same traits usually show positive correlations. These become higher as the appraisals become more similar in procedure and more complete in sampling, and as the groups of persons become larger and more similar in environment. The actual correlations between self-ratings on questionnaires and behavior records are usually low. The correlations between ratings by teachers, counselors, and interviewers with behavior records are often low. The correlations between conduct tests and time-sampling methods are usually moderate. Since all of these appraisals have been shown to have high self-consistency, one must conclude that they are evaluating different aspects of the rater, the ratee, or both. A more analytical approach is needed to discover basic patterns of behavior. Such an approach is given in part by correlational analyses.

II. CORRELATIONAL ANALYSES

In view of the fact that nearly all of the appraisals of modes of behavior thus far discussed secure total scores by adding together items which are somewhat ambiguous, factorial analysis of scores cannot be expected to give very clear results. Such analyses, however, are of value in showing whether the items which have been grouped together on the basis of logical or empirical considerations are found by statistical analysis to depend upon one or upon several factors. The statistical approach may, if carefully interpreted, furnish an answer to the question, "What are the main independent modes of behavior in a particular group of persons?" The methods of Thurstone and of Spearman are most frequently used. Since they employ somewhat different procedures, both will be illustrated.

A. Results from Thurstone's Method

Vernon (1938) applied Thurstone's technique to the results of the elaborate Boyd Personality Questionnaire given to 50 men and 50 women student teachers. The 175 correlations among 19 separate scores averaged .366 ± .06 when corrected for attenuation. Only 4 independent factors were needed to account for the correlation matrix. Vernon identified them tentatively as:

1. *Self-depreciation:* This was prominent in items which emphasize depression, instability, anxiety, shrinking from responsibility, and lack of self-sufficiency. This was the largest factor. It accounted for 41 per cent of the total variance. The next 3 factors together accounted for only 35 per cent.
2. *Carefreeness:* This was prominent in items representing suggestibility, freedom from worries, dissociation, inability to concentrate, lack of definite interests, and freedom from tenseness.
3. *Scrupulousness:* This is found in obsessional carefulness, strong self-control of feelings, freedom from emotional thinking, strong concentration, and acting readily without pressure.
4. *A factor which differentiated men from women:* Women showed stronger dislikes and fears, and more instability and dependency. Men showed more scrupulousness, inability to concentrate, and introspectiveness.

A careful study by Layman (1937) obtained 12 factors from correlations between 67 items which had been answered by 276 students. The factors were tentatively named:

1. *Gregariousness*
2. *Social Inadequacy*
3. *Social Initiative*
4. *Social Aggressiveness*
5. *Changeability of Interests*
6. *Self-sufficiency,* or *Independence*

7. *Inferiority, or Lack of Confidence*
8. *Impulsiveness*
9. *Moodiness*
10. *Sensitivity*
11. *Emotional Introversion*
12. *Inability to Face Reality*

The first four factors in this list are modes of social adjustment. They indicate four independent modes of behavior, not a single elemental sociability. Although these four modes are not easy to distinguish, they seem to be reasonable subdivisions of social activity. Thus, a person may be gregarious — that is, desire the company of others — because he feels socially inadequate, because he is aggressive, or because he desires to initiate some co-operative venture. There may be other independent aspects of social contact which will appear in other investigations. The last eight factors listed by Layman seem to be emotional adjustments. It is difficult to secure an accurate verbal description of emotional modes of behavior from this material, but the factor names are fairly descriptive and they confirm the work of others.

Another thorough investigation of traits was made by Mosier (1937), who followed a procedure reported by Guilford (1936). Mosier selected from several lists of personality traits the 42 items which seemed most diagnostic of individual differences. These 42 items, which are shown in Table 94, page 539, were administered twice with one week intervening to 500 male students at the University of Florida. The scores for the 42 items and also for the American Council on Education Psychological Examination were intercorrelated. From the matrix of 780 correlations, eight independent factors emerged through the use of Thurstone's Center of Gravity method. These were reduced to the simple structure shown in Table 106 by graphic rotation. The highest loadings in each factor are underlined. The tentative interpretation of these factors is given on the basis of those items which show the highest loadings:

Primary Trait Matrix

For the Items Listed in Table 94, page 539

Item	*Loadings on Primary Traits*								
No.	*C*	*D*	*H*	*I*	*S*	*P*	*Co*	*Au*	h^2
1.	048*	026	131	252	234	766	017	– 098	738
2.	– 079	118	075	048	572	350	– 235	– 045	532
3.	250	020	380	067	302	– 117	064	183	355
4.	073	679	006	181	026	– 032	– 032	467	719
5.	206	129	452	– 156	163	121	– 034	199	363
6.	251	115	518	077	065	032	112	119	378
7.	– 057	282	– 068	001	668	128	– 043	– 067	563
8.	026	200	260	– 089	155	– 055	042	467	363
9.	298	271	236	– 011	– 010	– 076	050	481	451
10.	409	151	083	– 103	248	– 046	113	316	377
11.	808	141	– 011	010	– 057	013	– 012	274	757
12.	– 045	009	003	218	708	362	– 052	002	687
13.	023	018	106	221	– 054	000	075	620	447
15.	202	098	397	498	154	012	254	075	545
16.	289	– 125	240	221	000	– 106	179	168	279
18.	001	– 013	206	003	475	206	274	020	382
19.	010	097	471	307	– 057	016	244	046	386
20.	002	500	067	299	420	126	– 361	– 228	718
21.	142	131	080	170	479	139	– 013	461	531
22.	262	– 011	– 001	349	– 053	– 031	248	190	296
23.	165	066	008	– 090	104	203	549	066	386
24.	158	200	143	142	094	102	346	176	273
25.	290	001	417	– 004	– 134	174	– 006	342	419
26.	387	346	310	187	– 026	052	– 036	199	445
27.	– 026	033	– 009	180	278	720	149	037	656
28.	431	560	298	– 092	099	– 069	158	190	673
29.	252	033	047	– 002	063	– 057	507	403	494
30.	– 140	084	020	096	319	503	328	– 018	490
31.	513	591	200	085	116	016	181	142	698
32.	161	714	088	035	161	– 045	100	490	795
33.	045	– 092	080	293	459	419	– 018	223	534
34.	100	236	046	655	499	172	– 018	– 030	766
35.	005	040	– 022	282	184	799	– 036	– 036	760
36.	– 003	022	023	109	344	184	– 019	475	388
38.	147	081	067	– 032	214	– 041	267	– 038	153
39.	145	042	041	425	595	113	045	209	619
40.	317	053	256	144	306	106	196	180	364
41.	849	199	038	037	014	– 128	– 035	246	851
42.	– 033	066	– 134	585	516	049	004	– 018	635
43.	017	– 019	228	174	083	– 101	– 428	172	308

* Decimal points properly preceding each entry have been omitted.

(Mosier, 1937, p. 283. By permission of the Editor, *Psychometrika*.)

1. Cycloid: The first trait is closely identified with the tendency to have wide mood swings.
2. Depression: This is shown by feelings of loneliness, sadness, depreciation, and anxiety.
3. Hypersensitivity: This trait is psychological rather than physiological. It is represented by hurt feelings and an inability to stand criticism.
4. Inferiority: This is shown by lack of self-confidence both in social situations and mechanical and thought problems.
5. Social Introversion: This is typical of persons who are shy and self-conscious in small, informal groups.
6. Platform Shyness: This is associated with stage-fright and appearance among strangers.
7. Mental Ability, also called Cognitive Defect: This is shown by poor grades in school, and on the American Council Test.
8. Autistic Tendency: This is shown by frequent daydreaming and wishful thinking of an emotional sort.

Table 106 also illustrates how factorial analyses may be used to construct better scales by selecting items which are completely explained by only one factor. If an item shows heavy loading in more than one factor, its interpretation will be difficult. Thus, Item 11 is good in the sense that it depends upon only one factor, but Item 2 is less valuable for analytical purposes because it has moderate loadings in two variables.

The per cent of variance due to all the factors is shown in column h^2. Item 41 is good because its variance is nearly all accounted for in this analysis. It has very small random or specific loadings to reduce its value as an analytical tool. Specific factor loadings can probably be reduced still more by factor analyses of more elaborate scales. Item 16 is of little value for careful measurement because its h^2 shows that only 28 per cent of its variance is accounted for.

Mosier interprets his results to mean that each of the factors will probably not be found on subsequent research to contain elements found in any other factor. The writer believes, however, that each factor may, on subsequent analysis, appear to be a composite of several other more basic and elemental factors. The existence of these 8 independent factors is supported to a considerable extent by the work of Guilford (1936), Whisler (1934), Vernon (1938), Layman (1937), and several others.

Carter, Conrad, and Jones (1935) used Thurstone's method to analyze an inventory of children's annoyances and a measure of intelligence. They found three independent factors called (1) general annoyability, (2) annoyance at untidiness, and (3) personal annoyance from interest in self-esteem. Intelligence showed a negative relationship with the tendency to be annoyed.

A factorial analysis by Burt (1938) is of unusual interest because he arrived at essentially the same results using correlations among traits and among persons. His sample included 124 persons chosen from a larger group because they all had similar average ratings on eleven emotional tendencies. Analyses following the methods of Spearman, Thurstone, and Kelley produced factors of general emotionality, of aggressive-inhibitive emotions, and of pleasurable-unpleasurable emotions.

Thorndike (1936) applied Thurstone's (1935) method to social intelligence and mental alertness tests. He found that the social intelligence tests measure primarily the ability to understand and to work with words which is such a large factor in verbal intelligence tests.

A factorial analysis was made by Brogden (1940) of intelligence and character tests applied to 100 sixth grade children. He used the Otis Group Intelligence Test and thirty performance tests which included four designed to measure honesty; three, perseveration; three, persistence; two, slang usage; and one to indicate each of the following: inhibition, suggestibility, conscientiousness, deportment, and grades in school subjects. By using Thurstone's (1935) method seven factors were clearly delineated.

1. *Resistance to suggestion*, similar to Spearman's W
2. *Honesty,* as shown by unwillingness to cheat

3. *Persistence*, as shown by continuing work in spite of fatigue, boredom, or distraction
4. *Verbal facility*, as shown by the Otis Test, similar to Thurstone's *V*
5. *A factor probably related to reasoning*, not clearly identified
6. *Self control*, or *dutifulness*
7. *Acceptance of a moral code*

B. Spearman's Method

During Spearman's research on cognitive experience, he found evidences of traits which corresponded to variations in energy. These were called *perseveration, oscillation,* and *will.*

1. Perseveration (P) and Its Opposite, Fluency (F). The idea that persons differ in their ability to shift rapidly from one activity to another has been investigated by many since Wiersma (1906) reported tests of speed of sensory adaptation. A survey of elaborate studies by Lankes (1914), Cattell (1934), and Clare (1936) shows that five sorts of tests have been used:

a. Persistence of sensory after effect, as shown by
  1. Speed at which different colored sectors of a color disk fuse, called flicker limen
  2. Time needed for light and dark visual adaptation
  3. Time needed for recovery of hearing after a loud noise
  4. Time needed for recovery of touch after a severe electric shock

b. Spontaneous recurrence of an experience
  1. In free association the tendency to give the same reaction to the same or to different stimulus words
  2. Direct questions about tunes, poetry, phrases, problems, and dreams coming to mind again and again

c. Hindrance of new mental activity by similar past activity
  1. Comparison of writing S's continuously with writing them as they appear in a mirror. This reversal technique is also applied to various letters, numbers, and drawings
  2. Comparison of immediate recall of a drawing with recall which has been delayed by the exposure of a second drawing. This technique was also applied to short narratives

    3. Direct questions on effects of being interrupted in various tasks, homesickness, seasickness, desire for change, and tendency to finish a task, although a reason no longer exists for completing it

d. Usual rates of activity

    1. Natural rate of tapping. The subject was told to tap with his finger just as he feels inclined at the time
    2. Speed of free association

e. Emotional patterns

    1. Aroused with difficulty but long in duration
    2. Pessimism and lethargy
    3. Few likes
    4. Either unusually submissive or else negativistic
    5. Inability to make small decisions quickly
    6. Either untruthful or punctiliously truthful
    7. Gives up easily

The results of applications of such tests, usually to small numbers of students, have been summarized by Spearman (1938). From correlation analyses he concludes that there exists in various amounts in each person a tendency for mental processes to lag. It is measured by inertia or slowness in shifting energy from one arrangement to another.

Line and Griffin (1935) applied a factorial analysis to tests of word association, reaction time, oscillation, perseveration, and Bernreuter scores, in an attempt to find the factors underlying mental health. A major factor emerged which was called "objectivity" and thought to be related to Spearman's fluency (F), which is the reverse of perseveration or inertia (P).

2. OSCILLATION (O). Another energy adjustment which Spearman finds unique is fluctuation or oscillation. It is supposed to be shown by:

a. The duration of attention shown in the waxing and waning of faint sensory stimuli, such as a light or small weight

b. The fluctuations in stereoscopic vision between the patterns which will not fuse, and also in monocular vision, the fluctuations in reversible perspective drawings

c. Changes in rate of continuous work in aiming or cancellation of test, or crossing out circles

Strangely, measures of this factor are not related to perseveration. Although both measure rapidity of change of some central mechanism, oscillation is dependent upon recuperation from fatigue, while perseveration is not due to fatigue but to difficulties in changing the direction of energy. By analogy, oscillation would correspond to variations of the steam pressure in a power plant; perseveration to the time required for the steam to go from the boiler to the cylinders; and *g*, to the average steam pressure maintained in the plant.

3. WILL (W). On the basis of quantitative results, Aveling (1926) concluded that conation in the sense of striving for a goal was quite different from volition in the sense of resolving or selecting a goal. If we accept the former as a rough definition of will, it is possible to assemble three types of appraisals which have been used by many investigators:

a. Tests involving variations in effort, shown by test of:
  1. Quality of handwriting, Courtis (1925)
  2. General mental tests
  3. Rate tests
  4. Tests of persistence in the face of distractions (see Hartshorne and May, 1930)

b. Ratings of behavior, such as those by Webb (1915) and Bernreuter (1931)

c. Observations of effort and persistence, such as time-sampling and log records

Results from variations of incentives indicate that the amount of effort which is effective usually increases with the complexity of the task or the speed with which the task must be performed. Very great effort, however, such as produced by a much desired prize, tends to increase speed at the expense of accuracy, even when aimed at producing more accuracy. Effort, when effective in unspeeded tests of complex sorts, seems to direct one toward relevant processes and continued action. In group measures variations in effort seldom produce any significant changes in intercorrelations.

Maller (1934) reported an analysis of correlations of measures of four aspects of character, called Honesty, Co-operation, Inhibition, and Persistence. To measure and estimate such behavior he applied an elaborate schedule of tests used by Hartshorne and May. A total of 708 pupils in three schools which served persons in upper, middle, and lower economic levels, were tested. The correlation between the scores taken to represent the four phases of character were low but positive, mean .29. The tetrad differences resulting from the analysis of separate tables of correlations for each school group were extremely small and in no instance more than three times their respective PE's. According to Spearman's logic, all the correlations may, therefore, be attributable to the presence of one common factor. When two measures of mental ability were introduced into the correlation matrices, the tetrad differences were very large; hence the factor common to the behavior tests cannot be identified with *g*. Maller believed that the common factor was a readiness to forego an immediate goal for the sake of a remote but more valuable goal. Behavior of this sort was demanded in all of the tests, and was typical of the will (W) factor described by Spearman. Tests of honesty and co-operation doubtless involve in addition to a factor of will, other independent factors determined by moral ideals. In some cases it may take more effort to be dishonest than honest. The intercorrelations among tests of honesty were usually low.

These analyses use the words will, effort, steadfastness, determination, and persistence somewhat interchangeably, and leave one with a vague impression of the meaning of the mathematical factor (W). Spearman (1938) also fails to be specific on this point. Apparently, "will," defined as effort, may refer to a general discharge of energy in many activities including both cognitive and muscular processes.

If the discharge is too feeble or too violent, it reduces *g*, the mental process which demands complex comparison. "Will" may also refer to persistence in the face of distractions implying a channelizing of energy to increase success in a particular situation which may or may not require much of the *g* factor. As used by Webb and by Hartshorne and May "will" is identified with kindness, trustworthiness, and co-operation, which are social goals rather than effort or distribution of energy. More careful definition and research is needed to clarify this factor. It will probably be resolved into two or more subdivisions with further scrutiny.

C. Conclusions

It is too early to write an adequate summary of primary traits. The field is being very actively investigated. Some general conclusions, however, are in order.

1. The results of factorial analysis depend upon the variety of items, the ambiguity of the items, and the variety of the persons who respond to the items. When few persons (less than 300), and few items (less than 40) are used, the investigation suffers from a paucity of facts. The best results will doubtless come from factorial analyses of unambiguous items, rather than from analyses of total scores from several divisions of a test. Total scores invite ambiguity through the arbitrary selection by the author of items which seem to indicate a particular trait. When items which are really independent of one another are added together, no quantitative analysis is feasible.
2. An inspection of the illustrations above shows rather startling similarities between results of measures on different populations. Spearman's *g* factor, and its variations, inertia (P) and oscillation (O), are found in all other investigations where comparison and energy factors are evaluated. Until the similarities have been pointed out,

these factors are usually given other names by other investigators.

A group of factors underlying particular moods and mental disorganization are well described by Mosier's work. One of his factors is quite like Spearman's *g*, but the other seven are patterns of ineffective modes of adjustments which seem to be dependent upon social experiences and physique rather than intellectual activities.

It is interesting to note that none of the factorial analyses have reported traits which closely resemble the mechanisms described by many psychoanalysts, such as projection, introjection, and fixation. This writer believes that these results do not preclude the existence of such mechanisms but illustrate the effects of selection of items on the resulting factors. The items in the available inventories have usually referred to characteristic moods of specific reactions, not to methods of resolving conflicts. If items of the latter sort were included in large numbers, a correlation analysis would almost certainly show factors corresponding to mechanisms. This would be a very interesting and valuable addition to the present inventories. Results of Rorschach tests have not yet been subjected to quantitative analysis.

III. EVALUATION OF ITEMS

The need of careful evaluation of items in the field of trait measurement should be apparent since wide variations in scores have been reported from tests which seem very nearly alike. Eventually a good measuring instrument must show for each item its chance factors, its tendency to measure only one trait, its ambiguity, and its emotional effect. Chance factors have been discussed in Chapter 6. The selection of items which measure only one trait was shown above in the discussion of Mosier's correlation analysis. It remains, therefore, to discuss ambiguity and emotional effects.

A. Elimination of Ambiguities

In a written examination, ambiguity may be detected by asking persons how they interpret the items. This information may lead to a revision of items so that interpretation in more than one way is difficult. An excellent example is given by Benton (1935), who applied the Personal Inquiry Form of Landis and Zubin to 20 normal adults and 20 psychotic patients, and then asked each one to explain his answers. From 44 questions, he obtained approximately 100 interpretations. Of the 44 questions, 20 displayed qualitative differences, 17 showed quantitative differences, and 7 had no differences. A qualitative difference is seen in the interpretation of "Have you ever felt that life is a dream?" Some thought that this meant, "Life is very unreal – like a dream." Others, "Life is very easy and beautiful." A quantitative difference in interpretation involved ambiguity in the amount of a trait or quality.

The various interpretations were cast into another form of 100 items called the Interpretation Questionnaire. Both the Personal Inquiry Form and the Interpretation Questionnaire were applied to 90 normal and 100 psychotic individuals twice at intervals of from 3 to 21 days. Answers on the Interpretation Questionnaire and on the Personal Inquiry Form were tabulated to show differences in the amounts of variation in interpretation. The differences were most apparent among items which had been changed to reduce the quantitative variations. Thus, the original item, "Do you feel mentally inferior to your friends?" was changed to, "Do you feel mentally inferior to most of your friends?" The original item is quantitatively more vague than the revised item. The original item did not differentiate between normal and psychotic groups, but the revised items showed large differences. This study shows that the ambiguity of items was greatly reduced by changes to reduce the number of possible interpretations. The revised items proved to be more reliable on a retest and more discriminative of psychotic tendencies than the original items from which they were derived.

B. Emotional Reactions to Items

In any sort of test many persons are on the defensive, and some, on the other hand, are influenced by expansive desires to show off. Test situations or interviews which make a person self-conscious are therefore likely to affect scores differently than situations where one is at ease or unaware of observation. Many investigators have emphasized the superiority of indirect measures of motives and adjustment tendencies. Maller (1932) reported that although nearly 70 per cent of pupils felt somewhat embarrassed by direct questions about their own shortcomings, only 43 per cent were irritated by impersonal questions. He therefore preferred the character sketch type of test in which one identified oneself with a pen portrait. Rundquist and Sletto (1936) preferred to state their items in the third person rather than the first. Murphy *et al.* (1937) and Watson (1925) have concluded that any self-rating methods have little value, except for comparison with observations of trained workers and with tests of conduct.

Another indication of emotional response is seen in the responses to positive and negative items. Lorge (1935) reported that people showed a tendency to check either positively or negatively any kind of question on the Bernreuter. This tendency was greater in some persons than their consistency in checking one item *Yes,* and its opposite *No.*

Smith (1932) found that positive and negative items which were designed to be just opposites of each other did not call out opposite responses. Thus, "Feels he has failed at most everything he tried," was not statistically opposite to "Feels he has succeeded in most everything he tried." Smith believed that the items called out different emotions.

Rundquist and Sletto (1936) found that their negative or gloomy statements were more discriminative of morale than their positive or happy statements. The positive statements tended to be answered more uniformly than the negative, and also to have smaller correlations with total scores in all except the scale for family adjustments.

These studies indicate clearly that reactions to specific items vary in ways which can be measured, and suggest the possibility of appraisals of emotional behavior by comparison of responses to various items.

The question is often raised as to whether a rating is more truthful when the raters are anonymous. Spencer (1938) found that students in a fairly large group admitted that they would have answered some questions untruthfully had they been required to sign their names. These students were those who showed the greatest *mental conflict* scores; hence Spencer concludes that signing one's name materially reduces the effectiveness of the selection of those who need counseling help. Olson (1936) and Moore (1937) also found that frankness increases with the degree of anonymity of the report.

According to Johnson (1934) administration should take cognizance of the mood of the subject. He reported marked differences in *dominance* scores when the same subjects were feeling cheerful and while depressed.

Hartshorne and May (1929) found a large increase in names given for co-operative children by classmates when the raters' names were placed in the ballots, as compared with anonymous ballots, but not much change in names of non-co-operative children. When raters' names were required, children centered their votes on positive votes and fewer names, and also voted less often for themselves.

Part III

PERSISTENT PROBLEMS

20

EFFECTS OF PRACTICE ON TEST SCORES

I. SIGNIFICANCE OF PRACTICE

Nearly all scores of tests of behavior show some improvement with repetition. This fact is significant for both theoretical and applied psychology. Theoretically, the careful evaluation of qualitative changes may contribute much-needed information on the nature of learning, and help in a systematic explanation of behavior. In the applied field, the prediction of success in educational and vocational activities would be greatly improved if the individuals to be compared had the same amount of practice. For predicting future developments, it is seldom useful to compare two persons' scores directly, when it is known that one of them has had considerably more practice than the other. For instance, if it appears that persons usually improve 30 per cent on the second trial of a particular test, it would be useless to compare one person's score on a second trial directly with that of another who was taking the test for the first time. Although in actual situations it is rarely possible to evaluate previous practice or experience which contributes to success in a test situation, it is possible to find out how much various sorts of practice do affect the scores on certain tests when conditions are well controlled.

The words learning, experience, use, exercise, and practice all have similar meanings. No attempt has been made here to give them technical limitations. They all refer to changes in patterns of behavior which are due to a person's past experience, with the exception of fatigue, illness, and other physiological conditions.

II. CONDITIONS FOR MEASUREMENT OF PRACTICE

The measurement of practice effects is a good deal more complex than those who have never attempted it might assume. To compare the improvement of two persons on the same test, conditions should be arranged so that they have (*a*) equal amounts of practice, (*b*) equal motivation, (*c*) constant behavior patterns, (*d*) similar types of scores, (*e*) equal units of measurement, and (*f*) true zeros. These requirements will first be briefly described and later, illustrated from reported studies.

A. Equal Amounts of Practice

Some investigators define equal amounts of practice as equal time spent in practice with equal time between practice periods. Thus, if all persons in a group had 10 five-minute practice periods in subtraction, with five-minute intermissions, they would be said to have equal amounts of practice. Because of its convenience, many investigators have used this definition of equality, although they point out that it does not mean equal amounts of work. The quickest subjects generally finish two or three times as many items as the slowest in any rate test. Consequently, the fastest workers get more of certain kinds of practice than the slowest ones.

Other investigators have defined equal amounts of practice as equal items completed with equal accuracy, regardless of the time needed. In this case all persons in a group are said to have equal practice in subtraction when they have all done the same problems with equal accuracy. This procedure is rarely followed because it involves individual testing, but it has been used in a number of interesting studies. Either procedure is useful, but the results, of course, will tend to be different.

B. Equal Motivation

One of the most serious problems in the measurement of practice effects is the evaluation of motives. If one person is very competitive and aggressive, he will usually show much greater improvement than another who is indifferent. Changes in motivation also appear when a person becomes bored with a long practice series or distracted by fears, or by desires for other sorts of activity. Such variations in motives undoubtedly cause marked variation in practice curves. Most investigators have used fairly large groups and assumed that interest would remain fairly constant for the average. Others have tried to secure ratings of interest from both the subjects being tested and other observers. Ratings generally give results which help to explain the most unusual cases, but they fail to show the effects of moderate changes in motives. A few experimenters have varied the incentive by offering different rewards. Studies of incentives, reported in Chapter 15, are interesting because they show that the variations in scores due to changes in motivation are sometimes larger than the variations due to a long practice series.

C. Constant Behavior Patterns

A fundamental requirement for the measurement of practice effects is that each score shall represent the same pattern of behavior. This is a persistent requirement of all good measurement work. Obviously, if a person's score on the first trial of an assembly test depends on manipulation and reasoning, but his subsequent scores depend upon his memory of his solution of the first trial, then the behavior pattern has radically changed. In order to have scores on an assembly test represent similar abilities on all trials, the memory factor should be eliminated or controlled so that all persons compared have approximately the same patterns of behavior. This requirement is a basic one for all careful comparison but it has been almost universally disregarded owing to the enormous difficulty of detecting and controlling changes in pattern. None of the studies described in this chapter have attempted such controls. Indeed, qualitative changes in performance are usually expected with practice. Nevertheless, there are ways of detecting and measuring the factors involved in success. These are discussed below where examples of factorial analyses applied to practice effects are described.

D. Similar Types of Scores

When time is measured in a test situation, it is often possible to express the results by two types of scores: the amount of work done per unit of time, or the time needed for a standard task. These are illustrated in Figure 69, which shows norms in the Minnesota Spatial Relations Test for both types of scores. The time-per-item scores decrease with practice and show smaller dispersions. The amount-per-time scores show just the opposite results. They increase with practice and show larger standard deviations. The two types of scores are usually interchangeable.

Time-per-item scores are preferred when a task needs to be finished in order to obtain a score. This is generally true in form boards, mechanical assembly tests, and a number of problem-solving tasks, where the end action is the only tangible result, since the preceding steps form a complex pattern of solution which cannot be easily observed. In some cases, steps toward the solution may be timed and scored accurately, but the total solution may involve abilities different from those required for separate steps.

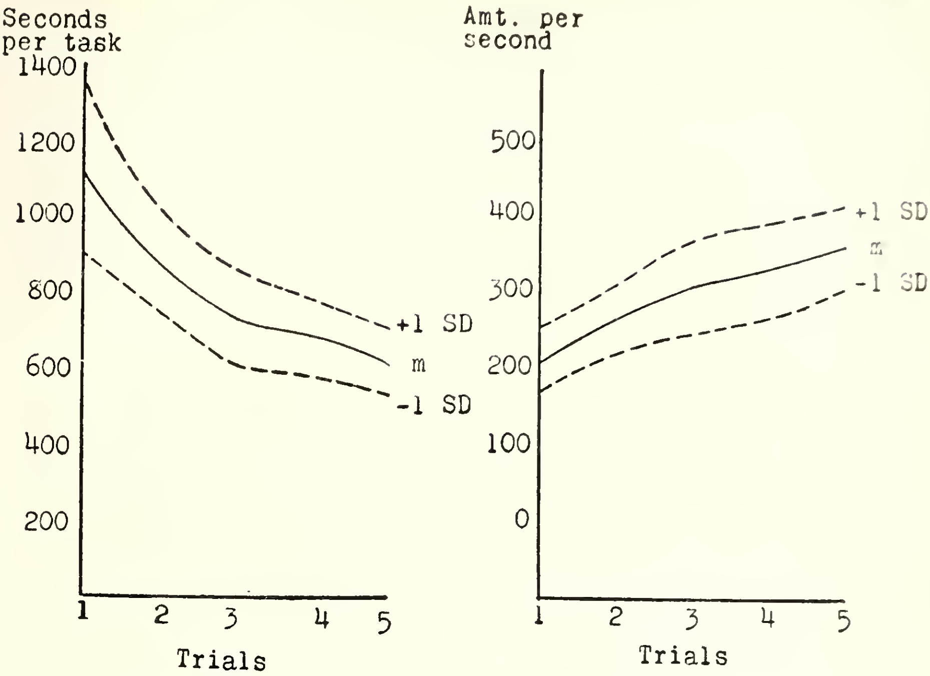


FIGURE 69. **Effects of Practice**

Means and Standard Deviations of Minnesota Spatial Relations Test, 58 College Students

(Greene, 1936. Adapted from Table 1. By permission of the Editor. *Occupations.*)

If the items are all similar and easy for a group, the amount-per-time is the only score available, as in rate of tapping or of addition. This situation is ideal for the use of amount-per-time scores. If the items are of varying difficulty, then an amount-per-time score is hard to interpret, as some items will take more time than others.

In tests where items vary a great deal in difficulty, such as the Otis Self-Administering Test, or the United States Army Alpha, two persons may get similar scores for different reasons. One person may be so cautious and methodical that he is just getting nicely started when told to stop. The other person may be so fast and superficial that he will have finished all he could in a short time allowance. One way to avoid such errors of measurement is to use only items of equivalent difficulty in a speed test. If samples of various levels of difficulty or complexity are desired, tests can be constructed for each level. The scores from such tests will show a person's ability at each level of difficulty. Tests which contain only one level of difficulty are by far the most accurate for the appraisal of practice effects, since the results have less ambiguous interpretations than tests where several levels of difficulty are used.

E. Equal Units of Measurement

In many practice situations, the items are of greater difficulty as one progresses. Thus, nearly all tests of sensory acuity and most of the general and achievement tests are progressively graded in difficulty. Such altitude scales often show more gains among the less able examinees who are working on the easier tasks than among the more able working on more difficult tasks. If scaled in equivalent units the opposite result might appear (see Chapter 22).

When a task is repetitive, such as a rate test of aiming, or cancelling all the *a*'s on a page, then the units of the scale are nearly equivalent in the sense that they may be substituted for one another. With practice such individual tasks may actually become easier, but an increase in scores requires an increase in speed of movement or adjustment between tasks. The total pattern, therefore, requires a more complex or delicate adjustment for the faster rates. In the case of practice in telegraphy or typing, a person is developing more complex underlying patterns of response, so that he may show little improvement in score for a time, then great subsequent improvement when the underlying patterns become effective. Such plateaus in practice curves are extremely hard to evaluate or predict. The problem of securing equal units in tests of this sort is a difficult one. The observer who works carefully will always try to show exactly the patterns of behavior which are represented by a score.

F. True Zeros of Development

The amount of improvement is often expressed as the per cent of the first trial. This is a valuable method of comparing persons or tests only when a true zero is known and used. In most testing situations, the scale zero is arbitrarily fixed at a point where no credit is secured on the test. This point may be considerably above a true development zero of abilities needed in the test. If a person increases his score from 2 to 4, he has gone up 100 per cent of his first trial. If, however, the true zero is 50 points below the arbitrary zero, he has really changed from 52 to 54, or only 3.8 per cent. On a percentage basis, other things being equal, less able persons will nearly always show greater improvement than the more able, since the lower scores are nearer the arbitrary zero.

There are two ways of avoiding this sort of error. One is to try to locate the true zero; the other is to measure improvement on a scale of standard deviations, which is free from raw score variations and zeros. Both procedures, which are described in Chapter 22, can be used only when the group tested is large and normally distributed.

Although these six conditions just described are desired for accurate measurement of practice, they are so hard to secure that few, if any, studies of practice have been completed under ideal conditions. The results found in the literature of this subject often lead to specific findings which appear contradictory because of the method of measurement. When uniform methods are used, the results of various studies become comparable.

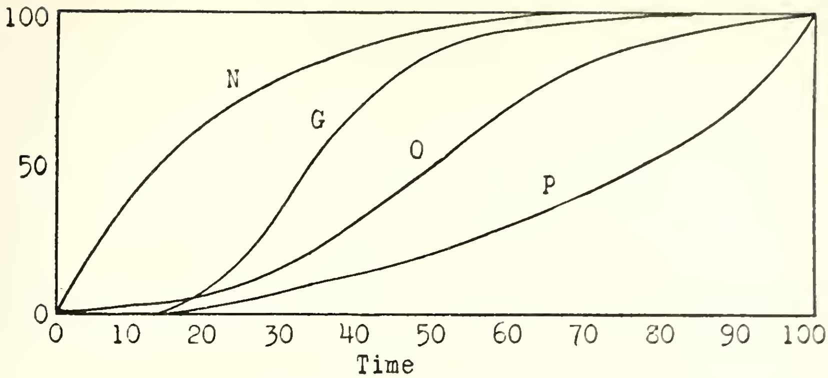
III. A UNIVERSAL PRACTICE CURVE

One of the important questions raised in any study of practice or learning is whether the shapes of all learning curves are essentially the same. If they are, then comparison of two persons on the same task would be relatively easy.

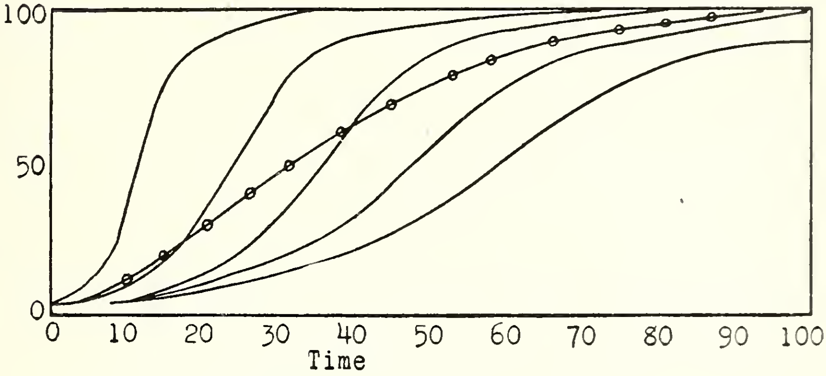
No conclusive answer can be made from data available at this time, but Thurstone (1930) and Courtis (1933) have both contributed theories to its solution. Both men emphasize the need for measures of individuals made under constant conditions of environment and motivation. Both conclude that, when these conditions are constant, the curves of learning for a given individual are regular and their later development predictable. Both point out the dangers of averaging the practice scores of several persons, as is the case in 90 per cent of the studies reported.

Figure 70 A shows four curves which illustrate possible practice effects. The curve marked N is negatively accelerated; that is, it rises by smaller amounts after each trial. Curve G shows small gains followed by larger gains, which, after an inflection point, are followed by decreasing gains per trial. Curve O is like Curve G except that the inflection point is just half way between the zero and maximum scores. Curve P is positively accelerated; that is, it shows larger increments per trial with practice.

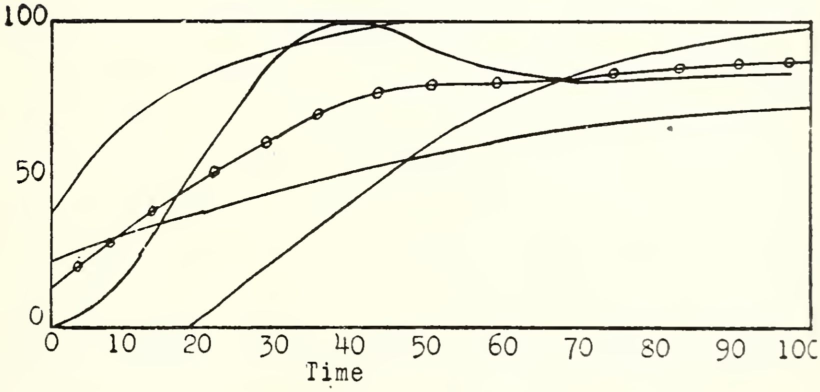
These curves are typical of results actually found in various situations, such as shown in Figure 71. There is usually some evidence that Curve N does not show the real beginning of practice and that Curve P does not show the end. If this is the case, then Curve N might represent the upper part of Curves G or O, and Curve P their lower extremities. This leaves Curves G and O for consideration. Thurstone believes that any practice curve of homogeneous material will have the shape of Curve O, which is the ogive of a normal frequency distribution. It is also made of two hyperbolas joined at the inflection point. He explains this as the summation of various ogive curves, as shown in Figure 70 B for five tasks of different complexities, which are equally weighted in the total score. Thurstone's (1930) discussion furnishes a clear rational approach to the measurement of learning functions, but does not give evidence that his basic assumption is correct — namely, that a curve



A. Four Typical Practice Curves



B. Summation of Practice Curves



C. Summation of Practice Curves Drawn to Represent a Common Situation

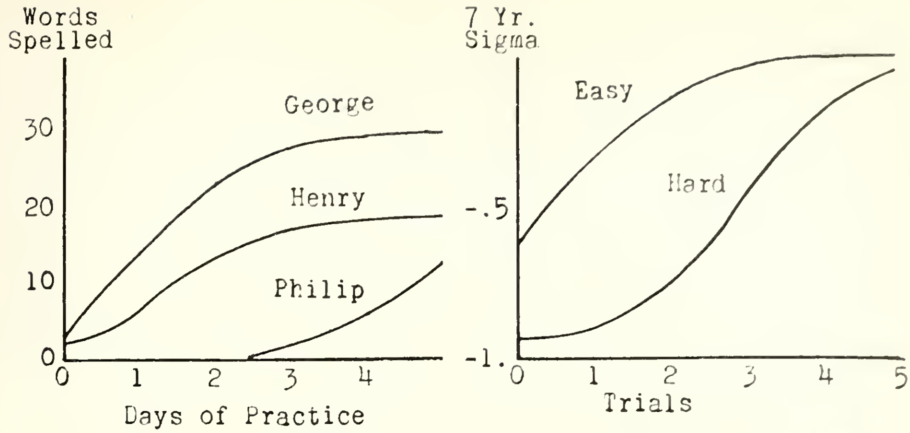
FIGURE 70. **Theoretical Practice Curves**

(B. From Thurstone, 1930, p. 481.)

for learning homogeneous elements under constant conditions will always be a symmetrical ogive. Empirical evidence is needed. Courtis (1925) believes that the results from the most complete and careful studies of practice all approximate Curve G with remarkable accuracy. He has collected a fairly large number of curves from various reports which certainly support this belief. Inspection of his report indicates that they may represent learning of complex or heterogeneous elements.

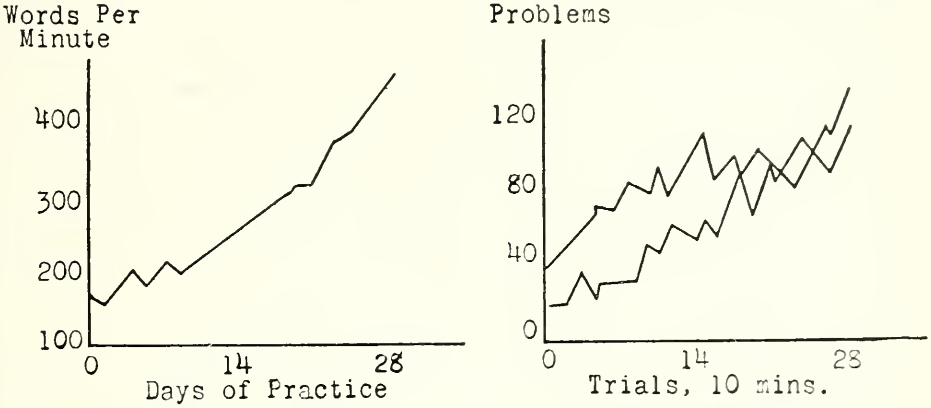
One of the obstacles in the way of discovering the true shape of practice curves is the fact that most of the practice curves reported are incomplete. They show neither the beginning nor the end of the learning of a skill, but only several points where the amounts of previous and subsequent development are unknown. These fragmentary studies do not usually afford enough evidence to verify a theory of a learning curve. Typical curves for practice on mechanical assembly, speed of reading, mental multiplication, spelling and maze solution are shown in Figure 71. They can all be made to fit portions of the curves marked G or O in Figure 70 A. All but two show negative acceleration toward maxima, which are apparently not reached. Inflection points are usually located somewhere in the first half of the period of practice. The hypotheses of either Thurstone or Courtis are supported by such data, but prediction to unknown parts of individual curves is very hazardous.

Another difficulty is the fact that many studies report changes with practice which introduce or eliminate factors which seem to have different starting points and maxima. The summation of such curves is shown in Figure 70 C, where four curves are averaged. This figure is probably more typical of practice situations than that in Figure 70 B. Without separate measures of each independent factor, no clear understanding of their combination is possible. This fact emphasizes the great need for tests which will measure only one independent pattern of behavior.



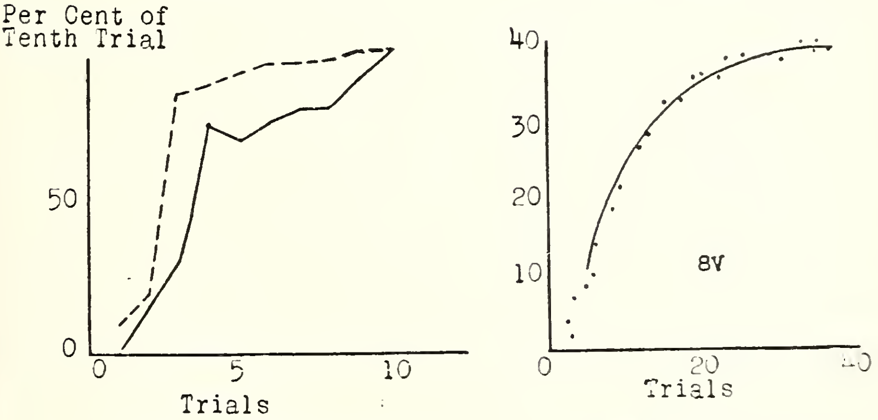
A. Practice in Spelling

B. Practice on Pencil Mazes



C. Silent Reading

D. Mental Multiplication



E. Stenquist Mechanical Assembly

F. Maze Learning Rats

FIGURE 71. **Typical Practice Curves**

(A. After Courtis, 1933; B. Greene, 1932, 7-year group; C. O'Brien, 1913; D. Brooks, 1929; E. Greene, 1937; F. Wiley, 1937.)

In spite of the difficulties of measurement, some evidence is available that practice under well-controlled conditions will be very accurately predicted. For instance, Wiley and Wiley (1937) showed that maze learning in rats followed a standard hyperbola curve with few deviations, Figure 71 F. After the first few trials, during which the animal made a number of adjustments to the situation, the results of practice were quite predictable for most of their animals.

IV. EFFECTS OF PRACTICE ON GROWTH CURVES

All practice curves are further complicated by the fact that they are superimposed on maturation curves, as shown in Figure 73 F. Here the results of a few daily trials are shown by a dotted line rising to a maximum, which, without the practice, would be reached only after several years. This illustration is typical of many curves for motor co-ordination, reasoning, and the learning of information. All practice curves must be thought of as starting somewhere along a maturation curve. Likewise, nearly all maturation curves of behavior must be thought of as being somewhat affected by practice effects. There is no way of separating these factors as independent components of a single score. Only when tests are repeated frequently under well-controlled conditions will the variance due to practice and individual maturation become clear. Two studies of the effects of practice on growth curves are reviewed.

Adkins (1937) found that pupils who took the Kuhlman-Anderson, Morgan, and Otis group tests of intelligence two or three times at intervals of a year had slightly higher scores than similar children who were taking the tests for the first time. She pointed to the need of correcting for such practice effects in longitudinal studies of growth.

Greene (1933) gave the Michigan Nonverbal Series four times, with one-week intervals, to 3,940 persons who were selected at random as samples at each age from 5 to 20 years.

The raw score means, shown in Figure 72, indicate that the mean performance for tapping with the right hand was not changed very much in four trials. The increases were approximately 3 per cent or 4 per cent of the first trials. The increases during four trials were equal to one year's growth for older groups.

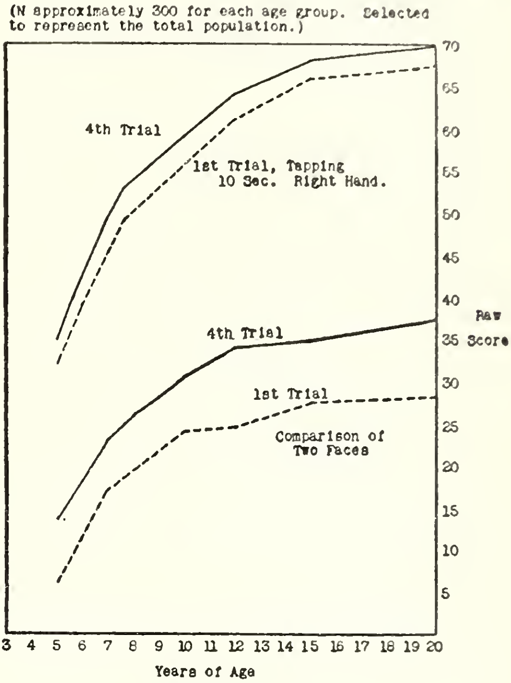


FIGURE 72

Effects of Practice on Growth Curves

(Greene, 1933)

In another test, Comparison of Two Faces, four trials improve the score approximately 20 per cent, or as much as eight years of growth thereafter. The improvements at age 5 and at ages beyond 10 were greater than for those between the ages 6 to 10; therefore, the growth curves have different shapes. The difference between the 5th and the 6th year is probably due to differences in training, since the 6-year-olds were in the first grade, where attention to printed detail is usually emphasized. The larger practice effect after 10 years of age is harder to explain. The experimenter thought it due to better motivation at the later ages.

V. MEASURED RESULTS OF PRACTICE ON TESTS

This discussion of the difficulties of measuring the results of practice should lead one to be rather cautious about broad generalizations. A number of interesting reports of the effects of repeating standardized tests several times will be presented under three headings:

1. Changes in mean scores
2. Changes in group dispersion
3. Correlations of trials

Since these three sorts of quantitative changes may occur independently, it is necessary to evaluate each separately. The qualitative changes which probably occur are discussed concurrently.

A. Changes in Mean Scores

1. GENERAL TESTS. A good deal of interest has centered around the possibility of increasing one's IQ on tests by practice or coaching. Experiments giving two trials during a short period are numerous and fairly conclusive. The Binet-type tests show small amounts of average improvement if repeated once within a few days. Terman and Merrill (1937) report that when Form L of the Revised Stanford-Binet was followed by Form M, or *vice versa,* there was a mean increase of approximately 2.5 points in IQ. Slightly larger increases were found below 5 and above 17 years. Although no one seems to have reported giving the Binet tests repeatedly day after day, tests given at 6-month intervals show indications of growth as well as practice. It is commonly reported by persons in clinical practice that several retests with the same Binet Scale within a few weeks are likely to raise scores by considerable amounts. Goodenough (1932) stated that the Minnesota Pre-school Scale should be corrected for practice to the extent of 2 C-score points or 3 IQ points, when repeated within a few weeks. A number of studies of specific coaching on Binet-type test, summarized in Chapter 10, show large and fairly permanent increases in total MA's and IQ's.

Studies of repeated trials of group-intelligence tests and point scales are also available. Thorndike (1922) gave fifteen equivalent forms of a mental alertness test, one each day, to college students. He found marked improvement up to the third trial and thereafter very slight improvement.

Rodger (1936) gave six group-intelligence tests to 76 children, 11 and 12 years of age. Their average IQ on the first test was 108.8, and this rose approximately 1 point per trial. The children with IQ's above 112 improved approximately 1½ points per trial; and those with IQ's below 80, ½ point per trial.

These reports show that improvement on general mental test scores, from the first trial to another within a few weeks, was usually from 3 per cent to 10 per cent of the first trial. Such reports do not indicate what parts of these tests are subject to the greatest improvement, nor the nature of the improvement. In some cases a skill is apparently developed; in others motivation is improved. In order to supplement these reports of general tests, practice effects on special tests are discussed.

2. SPECIAL TESTS. Only a few of the many studies available can be presented in a book of this size. Four reports are discussed below as typical of other studies on standard tests.

An article by Anastasi (1934) shows the effects of twenty repetitions by college students on four short tests: (*a*) cancellation of *a*'s in a page of fine type; (*b*) finding four-letter words hidden in a page of fine type; (*c*) symbol-digit and code learning; and (*d*) learning a new written language from a hard code. The results are summarized in Table 107. All the means increased from 50 to 70 per cent of the first trial, or from 2.8 to 3.8 SD's of the first trial. This indicates that the average person at the end of the training was surpassing the best person's scores at the beginning of training on these commonly used types of tests.

A study by Gates (1928) throws light on the effects of repeating tapping and digit-span tests over long periods. The rate of tapping with a blunt pencil was measured in a group

TABLE 107

Results of Practice of 114 College Students

	Scores in Trials				Increase*				
	First		*Last*		*Per*		*First*	*Last*	*r*
	M	*SD*	*M*	*SD*	*Cent*	*SD*	*CV*	*CV*	*FL*
Cancellation	40.6	6.8	59.7	7.9	48	2.8	16.8	13.2	.67
Hidden Words	43.6	6.9	69.3	11.4	61	3.8	16.0	16.0	.82
Code Learning	41.2	7.6	70.0	10.0	70	3.8	17.0	14.3	.30
Hard Code	39.1	6.8	59.3	8.9	51	3.0	17.4	15.0	.51

* Relative to first trial.

(Anastasi, 1934, p. 40. By permission of the Editor, *Journal of Educational Psychology*.)

of 82 children from 4 to 6 years of age, who were given three short practices daily for 18 days. At the end of this period the average had increased from 35.5 to 40.1 taps per period. The group was then divided by pairing the pupils for age, sex, mental age, IQ, grade in school, and scores on tapping and on eight other motor tests. One member of each pair was put into a practice group and given three short practice periods per day for 76 days, over a period of 6 months. The other member of each pair had no practice for 6 months, but then was given 17 days of practice. The mean scores of both groups at the end of the experiment were approximately 44.5 taps. The group that practiced regularly for 6 months improved only as much as the group which practiced for only 17 days before and after this period.

A similar group of children was given individual practice in recalling series of digits which varied in length on each of 78 days during a 5-month period. At the end of this time the average had increased from 4.33 to 6.40 digits, or approximately 6 years of mental growth according to the Stanford-Binet norms. This increase amounts to 47 per cent of the first trial. Four and one-half months after the last test, a retest showed a drop to 4.73. After 22 days more of practice the average rose to 5.73. An equated group without previous training showed a mean rise from 4.83 to 5.92 digits in 22 days of practice.

Gates pointed out that the 5 months of practice did not result in any marked advantage in relearning the technique. He believed that both the tapping and digit-span results indicate that the nature of improvement in this experiment was a temporary skill of adjusting to the test situations and organizing elementary patterns of behavior that had already matured. There is no evidence here that maturation was speeded up by 5 or 6 months of practice, as the improvements were greatest during the first few days, loss of skill occurred when practice ceased, and relearning was no faster for the practice group than for an unpracticed group.

Jersild and Bienstock (1935) reported that approximately 20 practice periods of about ten minutes each, distributed over an interval of 10 weeks, resulted in a mean improvement of almost 100 per cent in rhythm scores among five behavior-problem children, 28 to 45 months of age. The increases were thought due to improved co-operation and interest as distinguished from improved ability.

Another study of the same type showed that 14 children who had about 30 ten-minute training periods were not significantly better in rhythm than 9 similar children without practice, who were matched as controls. All of these children were classified as not having problems and gave good co-operation from the start.

Mosher (1928) gave 2 trials of six tests of reading comprehension and two tests of reading rate to 104 fourth-, fifth-, and sixth-grade pupils at intervals of one day between forms 1 and 2 of each test. The two rate-of-reading tests show nearly identical results, Table 108, namely, an increase of approximately 9 per cent of the first trial mean score, in this case .17 of the sigma. The comprehension test results are more variable. The greatest increase is 25 per cent. One test showed a mean drop of 8 per cent while three have no significant changes.

Greene (1937) gave 4 trials on each of fourteen standard tests at intervals of one day to equivalent groups of college

TABLE 108

Results of Practice on Reading Tests

	1st trial		*2nd trial*		*Increase* *		r_{12}
	M	*SD*	*M*	*SD*	*Per Cent*	*SD*	
A. COMPREHENSION							
Courtis	35.3	12.9	44.3	13.9	25	.70	.74
Haggerty	53.3	27.7	52.2	19.7	− 02	− .05	.83
Monroe	9.8	3.4	9.9	3.1	1	.03	.76
Thorndike-McCall	21.1	4.5	19.2	4.4	− 8	− .42	.75
Stanford	138.8	40.3	142.1	42.5	2	.08	.93
B. RATE							
Courtis	173.9	57.8	191.6	66.3	10	.17	.77
Monroe	103.7	46.0	112.3	40.1	8	.17	.71

* Relative to first trial.

(After Mosher, 1928, p. 273. By permission of the Editor, *Journal of Educational Psychology*.)

sophomores. Four equivalent forms of a vocabulary test, speed-of-reading test, and the Michigan Nonverbal Series were administered to 253 students. On the basis of these tests, six groups of 25 students each were selected having the same median and range as the whole group. Each group spent only ten hours on the project; hence it was not possible for any group to take all the tests, but the results from the different groups seem fairly comparable because of the manner of selecting the groups. Commonly used measures of sensory discrimination, processes using language and number symbols, visual, auditory, and kinesthetic perception, and motor co-ordination were included. These tests are listed below under four types of scores:

1. Amount recalled, without time limits: Tonal Memory from Seashore (1919), the Michigan Vocabulary Profile, and an auditory digit-span test (1919)
2. Liminal discrimination without time limits: pitch, time, consonance, and Seashore (1919)
3. Amount completed in limited time: tapping, aiming, feature comparison, pencil maze, speed-of-reading, and equation completion

4. Time required to complete a task: Kohs' Cube (1927), Stenquist (1923) Mechanical Assembly, and the Minnesota Spatial Relations Test, and Paterson (1930). (The fourth type of score was changed to the third by taking the reciprocal of the time score for each task, as in Figure 69.)

The standard directions for administering these tests were used, with the exception of the Kohs' Cube and the Stenquist Mechanical Assembly Tests, the time limits for which are so short that some persons could not complete all the items. In these cases, the actual time needed to complete each task was recorded for each person, and then converted into an amount-per-second score.

The items in tests of types 1 and 2 were progressively more difficult, whereas all the items in types 3 and 4 were nearly equivalent in difficulty. Thus, the score on the vocabulary test depends upon knowledge of words which are progressively more rare or complex; but the scores on an aiming, tapping, form board, or cube design depend in this study upon repeating the same or very similar tasks in the shortest possible time.

The complete analysis of such test situations to show the nature and the interrelationships of component processes will be the work of years of elaborate research. A tentative classification of independent processes on the basis of the qualitative reports of 76 students and 3 observers made during or immediately following the test periods is shown in Table 109. Table 109 divides the changes noted into four main classes of activity on the basis of their apparent independence: memory, comparison, movement, and persistence. *Memory* or recall is subdivided into five divisions which list patterns that appear in different test situations. These divisions are important because nearly all psychologists can agree upon them, and because they correspond remarkably with the amounts of changes in test scores. *Comparison* in these tests always involves both perception and recall as well as processes of deducing relationship. It includes four

TABLE 109

Subjective Analysis of Independent Processes

1. Memory:
    - *a.* Memory span, or immediate recall, such as digits
    - *b.* Specific preliminary information, as in the vocabulary test
    - *c.* General preliminary information or skill, as reading habits and methods of work
    - *d.* Specific information learned during the test, such as the way to put a bicycle bell together
    - *c.* General information or adjustments learned during the test, such as good methods of approach to a problem situation
2. Comparison:
    - *a.* Direct liminal comparison, as in pitch discrimination
    - *b.* Direct supraliminal, sometimes called simple perception, as in easy feature comparison
    - *c.* Eduction of easy relationships, as in easy reading and simple feature and maze tests
    - *d.* Eduction of hard relationships or solving puzzles, as in the Stenquist Mechanical Test
3. Movement:
    - *a.* Speed of movement, as in tapping
    - *b.* Accuracy or precision of movement, as in aiming
    - *c.* Rhythm or grouping of movements, as is found in tapping, aiming, and form boards
4. Persistence:
    - *a.* Avoidance of distractions
    - *b.* Control of energy, as in long or baffling tests

(Greene, 1937, p. 71. By permission of the Editor, *American Journal of Psychology*.)

types which can usually be distinguished on the basis of complexity. *Movement* is not easy to isolate from the two previous types of activity. It would, however, be used in describing patterns in which muscular elements seem to produce the main individual differences. *Persistence* refers to the continuation of the other three types of activities just described. It involves the direction of energy toward a goal in the face of various external and internal distractions. These four seem to be the main variables which corresponded to recorded changes with practice. There were undoubtedly others which were harder to observe and to record systematically.

Although each test probably employed all of these processes to some extent, the scores were generally reported to be due very largely to one or two types of processes. Further, the subjective analysis of practice effects on these tests showed that with improvement one pattern tended to become more dominant. Less trial and error and less shifting of method were found after practice.

The various qualitative changes which occurred during practice were found to correspond closely to the differing amounts of improvement on the various tests. Those tests which had nearly the same scores throughout seem to depend on processes which change very little with practice. Conversely, if a test score shows great improvement with practice, it is probable that somewhat independent processes are either eliminated or introduced. This was particularly noticeable where a remembered solution was substituted for a puzzle-solving activity, as in the Stenquist Mechanical Assembly Test.

An inspection of Table 110 shows that those tests whose mean scores increased from zero to 5 per cent included principally two types of processes, either speed of ballistic movement or liminal comparisons. Increases from 6 to 25 per cent were found in tests which depended principally upon specific preliminary information, or speed-of-precision movements, or direct comparison. Increases from 26 to 75 per cent were found in tests of easy comparisons. Increases from 76 to 200 per cent were found principally in tests where a generalized rule or method could be learned during the tests; and increases of more than 300 per cent were limited to the puzzle variety of test in which a solution or a partial solution was remembered and used in a subsequent trial. Inspection of Tables 107 and 108 reveals similar tendencies which are not strictly comparable because of the variations in number of trials and types of persons. The results show very clearly the need for attention to the problem of devising tests whose main factors will remain the same throughout long periods of practice.

TABLE 110

Effects of Practice on Fourteen Standard Tests of College Students

Tests	Time Limits	Scores on Trials				N	Relative Increases from 1st to 4th Trial		Coefficients of Variation		Change	r_{14}
		1st		4th			Per Cent of 1st	SD of 1st	1st	4th		
		M	SD	M	SD							
1. Increase from 0 to 5 Per Cent												
a. Time	None	72.5	16.2	75.3	11.3	53	3.8	.17	22.2	14.7	− 7.5	.84
b. Pitch	None	82.5	12.5	82.6	12.4	53	0.1	.08	15.3	15.2	− 0.1	.87
c. Consonance	None	60.4	11.2	62.4	13.7	53	3.3	.18	17.5	21.9	− 4.4	.65
d. Tonal Memory	None	75.1	18.7	75.2	18.6	53	0.1	.05	24.9	24.8	− 0.1	.65
2. Increases from 6 to 25 Per Cent												
Language												
a. Vocabulary	None	22.3	6.2	25.1	6.2	92	12.2	.457	28.2	24.8	− 3.6	.90
b. Speed-of-Reading, Mich.	10 min.	65.5	25.3	87.2	17.1	284	33.8	.86	38.8	19.4	− 19.4	.83
Numbers												
c. Equation Completion, Peterson	15 min.	30.3	21.0	87.0	43.1	46	192.0	2.67	68.5	49.2	− 19.3	.74
d. Digit Span, Auditory	None	75.2	12.1	91.9	15.4	46	22.4	1.38	16.1	16.8	− 0.7	.85
3. Increases from 26 to 75 Per Cent												
Pictures												
a. Feature Comp. 1	2 min.	35.2	11.7	56.7	11.75	235	61.5	1.83	33.2	20.9	− 13.7	.66
b. Feature Comp. 2	3 min.	21.4	10.2	54.2	14.2	235	157.0	3.20	47.3	26.2	− 21.1	.54
c. Feature Comp. 3	5 min.	12.2	6.7	18.7	7.5	235	54.2	.97	54.7	40.1	− 15.6	.73
Mazes												
d. Pencil Maze 1	2 min.	18.0	4.2	30.4	5.7	235	67.4	2.95	23.2	18.7	− 5.5	.78
e. Pencil Maze 2	3 min.	18.2	10.2	42.6	14.2	235	137.4	2.39	55.6	33.4	− 22.2	.57
f. Pencil Maze 3	5 min.	9.0	6.5	36.6	10.5	235	307.0	4.25	66.7	28.2	− 38.5	.46
g. Tapping, Pencil and Paper	10 sec.	67.1	18.3	70.4	10.2	235	4.9	.18	27.2	14.5	− 12.7	.91
h. Aiming Pencil at 1½ mm. Circles, Right Hand	1 min.	84.7	15.3	98.2	14.3	235	16.2	.88	18.2	14.6	− 3.6	.85
4. Increases from 76 to 200 Per Cent												
a. No. 1 Kohs Design	None	5.2	2.1	12.3	1.9	23	136.4	3.38	40.4	15.7	− 24.7	.47
b. No. 5 Kohs Design	None	2.7	0.7	5.7	1.4	23	111.1	4.28	25.8	24.5	− 1.3	.48
c. No. 10 Kohs Design	None	0.92	0.2	1.8	0.3	23	100.0	4.50	22.2	16.6	− 5.6	.60
d. No. 14 Kohs Design	None	0.8	0.2	1.4	0.3	23	75.0	3.00	25.5	21.4	− 4.1	.75
5. Increases of 201 Per Cent or More												
Mechanical Appliances (Stenquist)												
a. Cupboard Catch 1	None	0.6	0.3	2.8	0.7	19	350.0	7.33	50.0	25.0	− 25.0	.32
b. Bicycle Bell 5	None	0.1	0.1	0.9	0.1	19	520.0	6.00	81.0	18.7	− 62.3	.30
c. Push Button 8	None	1.2	0.7	5.7	1.4	19	375.0	6.00	38.4	24.6	− 33.8	.24
d. Mouse Trap 10	None	0.05	0.1	1.1	4.5	19	1900.0	32.50	42.0	41.0	− 1.0	.07
e. Form Board: Minn. Spatial Relation A B C D Together	None	207.0	41.0	31.0	62.0	58	59.9	3.01	19.8	18.7	− 1.1	.47

(Arranged partly from Greene, 1937, Table 1. By permission of the Editor, *American Journal of Psychology*.)

B. Changes in Group Dispersion

The effects of practice on individual differences have been studied by a fairly large number of persons, stimulated perhaps by Thorndike's (1908) hypothesis:

> Experiments in practice offer evidence concerning the relative importance of original nature and training in determining achievement. In so far as the differences among individuals in ability at the start of the experiment are due to differences in training, they should be reduced by further training given in equal measures to all individuals. If, on the contrary, in spite of equal training, the differences among individuals remain as large as ever, they are to be attributed to differences in original capacity.

Anastasi (1937) questions this hypothesis on the ground that differences in training alone might, and probably do, provide skills, attitudes, habits, or specific information which would make persons differ still more with subsequent equal periods of practice on certain skills.

The traditional approach has been to compare raw score dispersions of various trials of tests. Thorndike (1908) and Wells (1912) using arithmetic tests and Chapman (1914) using verbal analogies and opposites found that practice increased the standard deviations of the group. Whitley (1911) using weight discrimination, cancellation, and mental multiplication; Peterson (1917) using ball tossing; and Gates (1922) using tapping and arithmetic calculations, reported that practice reduced the individual differences. Barlow (1924) using card sorting, mental multiplication, and substitution tests concluded that the more complex tests showed greater group dispersion with practice, and that the simpler tests showed less variability after practice. Peterson and Barlow (1928) reviewed 40 published studies and found that the coefficients of variation of 24 tests decreased, of 5 remained the same, and of 11 increased slightly with practice. Reed (1931) compared the coefficients of variation of 70 previously reported experiments. He found that

77 per cent decreased, 4 per cent showed no change, and 19 per cent increased with practice. Burns (1937) summarized 81 previous studies to show the types of motivation used. He found a serious lack of attention to motivation factors and believed that these must be carefully controlled to make the results comparable.

The data from the three reports, shown in Tables 107, 108, and 110, all show a marked tendency for the standard deviations of the first trials to be smaller than those of subsequent trials. A few exceptions do appear. Most of these exceptions are for tests of vocabulary or of reading comprehension which involves vocabulary of greater difficulty as one progresses through the test. Since these tests were not scored with equivalent unit scales, the failure of the standard deviations to increase with practice may be due to the scaling. Some of these tests did not take account of the time spent, so that some students may have spent twice as much time as others. This procedure would reduce the individual differences in many tests. Tables 107, 108, and 110 also show a marked tendency for the coefficients of variation to decrease with practice, indicating that the means usually increase faster than the sigmas on the raw-score scale. The few exceptions are in cases where no significant changes took place.

In a summarization of this discussion of the effects of practice on group dispersion, as shown by *amount* scores, it can be said that when any change occurred with practice the standard deviations of groups usually increased, while at the same time the coefficients of variation decreased. Results are occasionally reported, however, which show the opposite trends.

C. Correlation of Trials, Prediction

1. Logical Approach. A logical approach to possible combinations of practice curves will show what factors influence prediction. Let us assume that motivation, for

those to be compared, is identical and that the scores are recorded in terms of amount-per-time. Then the simplest practice situation would compare persons of the same age, environment, and having the same shaped curve of practice for a given test. The curves would be represented as in Figure 73 A. The similar environment means that individuals start to learn at the same age and have equal practice periods. All four curves in this figure rise proportionately with equal amounts of practice. Differences among individuals increase with each trial, and prediction from scores at one time to scores at a later time is perfect.

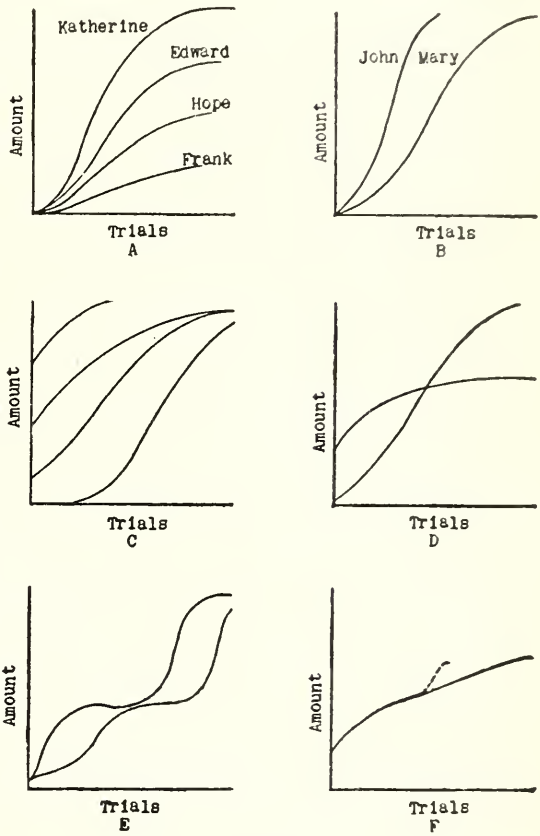


FIGURE 73

Theoretical Predictions of Practice

Figure 73 B shows a more usual and more complex situation. Here John and Mary rise to the same level of proficiency, or maximum, but Mary takes twice as long as John. Equal amounts of practice will at first lead to greater differences between scores, but after John has ceased to show much improvement, to smaller differences. This situation is found either where the rates of practice favor one child more than the other, or where one is more eager to learn than the other, or where their rates of maturation are different.

Another common occurrence is shown in Figure 73 C, in which all persons tend toward the same maximum even though they have had different amounts of practice before the first trial of the test, that is, different *starting points.* Here all scores tend to become more alike with practice, and the SD of the group becomes smaller. This graph is also typical of a situation in which all the children have the same maximum, or a situation where children have different maxima but the brighter children are not motivated to do their best, or when the test itself does not show the real improvement at the higher end owing to poor scaling or to the lack of more difficult items.

Another very common situation is seen in Figure 73 D, illustrating what happens when two persons with different maxima have different amounts of previous practice. Here, with equal subsequent practice their scores at first converge, then diverge. This is the situation when a youngster with medium ability has had a great deal more practice and encouragement than one with superior ability.

Still another possibility is seen in Figure 73 E, showing the case of two persons with similar maxima who have different curves of development. Equal practice will in this case cause the score to converge, diverge, then converge again, owing to plateaus in the curves. Plateaus or flat parts are often found in practice curves for complex material.

These illustrations show that there are four aspects of practice curves which must always be evaluated for careful comparisons of persons: (*a*) the shape of the curve, (*b*) the rate of development, (*c*) the maximum development, and (*d*) the starting point. Usually in group testing nothing is known about any of these aspects. It is easy to see why conclusive predictions are not forthcoming from such material.

2. Typical Results. A person's position in the group may change a great deal with practice. For purposes of prediction it is important to know how much change to

expect. Table 111 shows intercorrelations for the five trials of the Minnesota Spatial Relations Test illustrated in Figure 69. There are two clear-cut tendencies shown here. One is that the correlations between adjacent trials are closer than those between nonadjacent trials. The further apart the trials, the lower are the correlations. This indicates that individuals' positions in the group changed considerably from trial to trial. It does not show which persons went up or down or showed no improvement, but it probably means that they were developing at different rates from different starting points toward different maxima. These results for the Minnesota Spatial Relations Test are typical of nearly all group results that show any improvement with practice.

TABLE 111

Product Moment Correlations of Five Trials of Minnesota Spatial Relations Test (uncorrected), N = 58

Amounts-per-Second, Boards A, B, C, and D totaled

Trials	2	3	4	5
1	.748	.703	.658	.507
2	—	.807	.736	.555
3	—	—	.826	.772
4	—	—	—	.856

(From Greene, 1936, p. 2. By permission of the Editor, *Occupations, the Vocational Guidance Magazine.*)

Tables 107, 108, and 110 allow a comparison of the size of correlation between first and last trials. Comparisons among the tables are not warranted for in each study the number of students and number of trials are different. On the other hand, the comparisons within each table are significant, for the correlations refer to the same or similar groups. Table 107 shows that prediction in the Hidden Words Test was very much better (.82) than in tests of speed of visual perception and association of letters and words; for the other three tests show correlations between the first and the twentieth trials of .30, .51, and .67.

Table 108 shows correlations ranging from .71 to .93 between first and second trials among tests of reading speed and comprehension.

Table 110 shows correlations from .07 to .91 between the first and the fourth trials of small groups of college students on a variety of tests. Tests of liminal discrimination of sound, vocabulary, speed-of-reading, equation completion, auditory digit-span, tapping, aiming, gave correlations above .74. Tests of manipulation, observation, and inference, which allowed considerable reorganization of process, showed low correlations, particularly in tests either with very short or with no time limits. Both the short and the unlimited periods introduce sampling errors.

Correlations between initial scores and gains are ordinarily nearly zero or slightly negative. A negative relationship implies that those who are low on the first trial make larger gains than the rest. There are several possible explanations for these findings. One is that random errors of measurement contributed to this negative relationship; for if one has by chance been rated too high on the initial test, he will show a less than normal gain, and if his first score is below his true ability, he will tend to show a larger gain than otherwise. Another explanation applies when gains are recorded as per cents of the first trial. Then negative correlations between initial scores and gain may be due to the fact that a gain of two points represents a 50 per cent increase when the initial score is 4, but only a 20 per cent increase when the initial score is 8. Reed (1931) summarized 59 experiments by others over a period of 15 years. He reported that 93 per cent of these studies showed negative correlations between initial scores and gain, where gains were shown as per cents of the first trials. A third explanation is that the previous experiences of those tested have been different, and that those with the least experience have easier learning tasks than do those with the most experience. This is usually a reasonable assumption. It is much harder to gain an inch in the high jump when well practiced than when just beginning.

These considerations lead to the conclusion that a scale is needed which will allow gains to be measured in terms of units of the difficulty of improvement, rather than in terms of raw scores. Such a scale is provided by absolute scaling to some degree (see Chapter 22) and by the use of individual practice curves even more effectively. Courtis (1933) has described a method for comparing individual curves which offers an interesting approach. The account of his method, which should be consulted by interested persons, indicates that initial scores and gains with equal amounts of practice are always related in random fashion, and that all persons have equal gains with equal practice when measured in units based on their own curves of growth.

D. Relations between Measured Changes

One of the most significant findings which comes from a comparison of these measured results of practice is the evidence that all of the changes show marked relationship with one another when the optimum controls of measurement are maintained, or even partly maintained. These relationships appear most clearly from an inspection of Table 110.

There is a very noticeable relationship between two measures of relative improvement — namely, percentages of first trial mean scores, and standard deviations (SD) of first trial distributions. A mean improvement of 5 per cent or less corresponds to an improvement of less than .5 SD. Improvement of from 26 per cent to 75 per cent corresponds to changes of 1 to 1.5 SD units, and from 75 per cent to 200 per cent is equivalent to changes of 1.5 to 3 SD. Above 300 per cent represents the same changes as do 4 to 32 SD. SD scores were not used in these comparisons, for the groups were small and some of the distributions were skewed. When large, normally distributed groups are measured from true zeros, it is quite probable that the two measures of relative improvement will be interchangeable.

Table 110 also illustrates a marked relationship between mean score increases in standard deviations and decreases in coefficients of variation. These seem to be related to basic changes in behavior patterns. A large increase in mean scores is usually indicative of a large change in either speed or rearrangement of pattern. In nearly all of the test situations this change produces larger individual differences, but since the means increased faster than their standard deviations, the coefficients of variation are reduced toward approximately 14. According to Thurstone's (1928) logic, these differences in CV's are probably due to differences in arbitrarily chosen zeros. All CV's would be nearly the same if the true zeros were used. This hypothesis raises the extremely interesting hypothesis that nearly all mental skills will have similar CV's if measured carefully on large random populations. Much work is needed to check this hypothesis, but available results suggest that it is a probability (see Chapter 22) .

In Table 110 the relationships between correlations and the other measures of practice effects are not quite as marked, but they are noticeable. In general, the smaller mean score increases are associated with the higher retest correlations. Exceptions to this rule are fairly numerous, but are usually small in size. The fragmentary nature of the original data and the lack of adequate controls point to the need for more research. The presence of a low correlation probably indicates a radical change in procedure among many members of a group, but it may reflect only chance errors in measurement.

E. Factorial Analyses of Practice

Multiple-factor analyses have provided some interesting conclusions concerning the nature of practice effects. The most elaborate study is contained in three reports by Woodrow (1938, 1938b, 1939). In one study, 56 college students completed 39 trials, one a day, in each of seven tests, adding, digit-letter substitution, spot-pattern memory span, anagrams, cancellation, and visual estimation of length of segment of a rod. Raw scores for the first five days were pooled to furnish a stable initial score, and for the last three days for a final score. Individual gain scores were calculated and scores on 12 tests of the sorts included in intelligence tests were secured for each person. A factorial analysis of all of these items yielded nine factors. From a second study Woodrow reported results of 66 trials by 82 subjects on four rate tests: adding, digit-letter substitution, two-digit cancellation, and four-digit cancellation. The raw scores on initial and final days of practice and 21 other mental test scores were intercorrelated, and subjected to a factorial analysis. Several other groups were subjected to other fairly elaborate testing schedules. Woodrow's results may be summarized in part as follows:

1. Marked changes in the size of factor loadings occurred in all tests of the first experiment, and in the substitution test of the second experiment. The changes in all cases showed a marked reduction of the highest factor loading on the initial test. Usually there was a large increase in one other factor and small decreases or no significant change in all the rest. These results indicate that any factor which was most potent for success on the first trial always lost much of its value after a long practice series.
2. There was no factor which corresponds to gains. Gains correlated positively with the final scores and negatively or near zero with the initial scores.
3. Gains did not show any resemblance to IQ's scores or to a factor which resembled Spearman's *g*.
4. These results which were secured from raw scores might be radically changed if other scores, such as T scores were used.

McNemar (1936) tested 182 junior high school boys and found the relation of practice to two motor factors which appeared in a correlation analysis of five tests: Miles' Speed

Drill, Whipple Steadiness, Koerth Pursuit Rotor, Brown Spool Packing, and a Card Sorting Test. The general factor which appeared in all the tests was larger on the final than on the first trials. On the initial trials the general factor was largest in the Speed Drill and Spool Packing Tests, but on the final trials it was slightly greater on the Pursuit Rotor Test than on the others. On both initial and final trials a second factor was negative in the first three of the tests named above and positive in the last two. McNemar did not attempt to name this factor but it might be related to a speed of the movements needed in the Spool Packing and Card Sorting Tests.

With practice there was a marked increase in the correlations between the Steadiness and the Pursuit Rotor Tests; the Speed Drill and the Card Sorting Tests; the Steadiness and the Card Sorting Tests; the Pursuit Rotor and the Spool Packing Tests; and the Spool Packing and the Card Sorting Tests. The other intercorrelations showed little change with practice. The changes in correlations seem to have been due to changes in patterns of behavior in a test situation. McNemar pointed out that a further experimental attack would be needed to determine the presence of a general factor in tests of this sort and its significance in measurements of practice effects.

A factorial analysis by Greene (1939) reported the results of applying 12 nonverbal tests four times to a group of 394 fourteen-year-old boys. He used tests of aiming and tapping with right and left hands, three levels of difficulty in pencil mazes, and three in picture discrimination. Two analyses were made: one, of the first trials of all tests; and one, of the fourth trials. In each analysis five factors appeared and were rotated to give a simple structure. From inspection, the same five factors seemed to appear in both analyses. Tentatively, they have been named:

1. *Induction* (the isolation of unique visual patterns from a series of the disconnected patterns in feature discrimination)

2. *Trail finding* (the isolation of a visual path from the larger maze patterns)
3. *Speed of ballistic movements* (tapping)
4. *Speed of precision movements* (aiming)
5. *Ambidexterity* (left-hand aiming and tapping, and easy maze solution)

With one exception, the analysis of the first and the fourth trials showed different factor loadings for the various tests. The test of aiming at the larger circles (4 mm. with the right hand) showed approximately the same loadings throughout the practice series. The other tests all showed marked shifts toward a larger loading in one factor and smaller loadings in all the other factors. This study is interesting because it demonstrates the high independence of ballistic and precision movements, a factor called ambidexterity which appears in both the motor and maze tests of right-handed people, and two factors of comparison and observation. The study also reported very similar factor patterns for the hard and the easy levels of all the tests, i.e., aiming, maze solution, and feature discrimination. These results are somewhat different from those of Hertzman (1936), who reported evidence of different factors in hard and easy levels of numerical and spatial tests. Much research is needed on the question.

VI. AVOIDANCE OF PRACTICE EFFECTS

Practically, it seems clear that only those test scores which show from zero to 5 per cent improvement can be used for comparisons of individuals without a correction for practice. In practical situations it is desirable to avoid practice effects instead of trying to measure them and correct for them. Two procedures have been suggested. One involves the repetition of equivalent test forms until all persons have reached a high degree of efficiency or show no further tendency toward improvement. This method is extremely laborious and expensive and may not show desired results even under the most standardized laboratory conditions because motivation is hard to control. The number of repetitions necessary for each person will vary according to factors such as the nature of the task, the maturity of the individual, his inherited tendencies, and his motivation. On the other hand, a few practice periods are often used in measurement, in order to overcome initial inhibitions and to develop normal adjustments to the test situation.

Another procedure depends upon a hypothesis suggested by Courtis (1935) — namely, that a ratio between scores on two tests which vary in only a single factor will be constant throughout practice, and will adequately represent one's ability in the skill which is common to both tests, or in some aspect of that skill. Work on this hypothesis is now being carried on, and it should prove very fruitful.

VII. MEASURED RESULTS OF FORGETTING

Among studies which have used standardized tests, the effects of practice upon individual scores seem to have received more attention than the effects of various periods of forgetting. This emphasis is unfortunate because, in the absence of evidence, there may be a tendency to believe that the effects of training are more permanent than they actually are. Tests of immediate recall or recognition of verbal or pictorial materials have been discussed in Chapters 8 and 10. Most of the reports of forgetting over periods of an hour or more are based upon either laboratory studies using mazes or nonsense syllables, or school studies using tests of specific information.

Typical amounts of mean retention over various intervals are shown in Figure 74, which compares the work of Ebbinghaus, Austin, Greene, and Jones. The pioneer work of Ebbinghaus (1885) for the retention of nonsense syllables is represented in part by a curve which drops very rapidly from 100 per cent to 30 per cent recall in a few minutes, and to 25 per cent recall after 1 to 4 weeks. Austin (1921) had mature adults read over technical passages five times using distributed and undistributed repetitions. She found that the average student retained approximately 75 per cent of the information at the end of the learning periods. The spaced method gave better retention than the unspaced, particularly at the end of 4 weeks. Jones (1922) measured

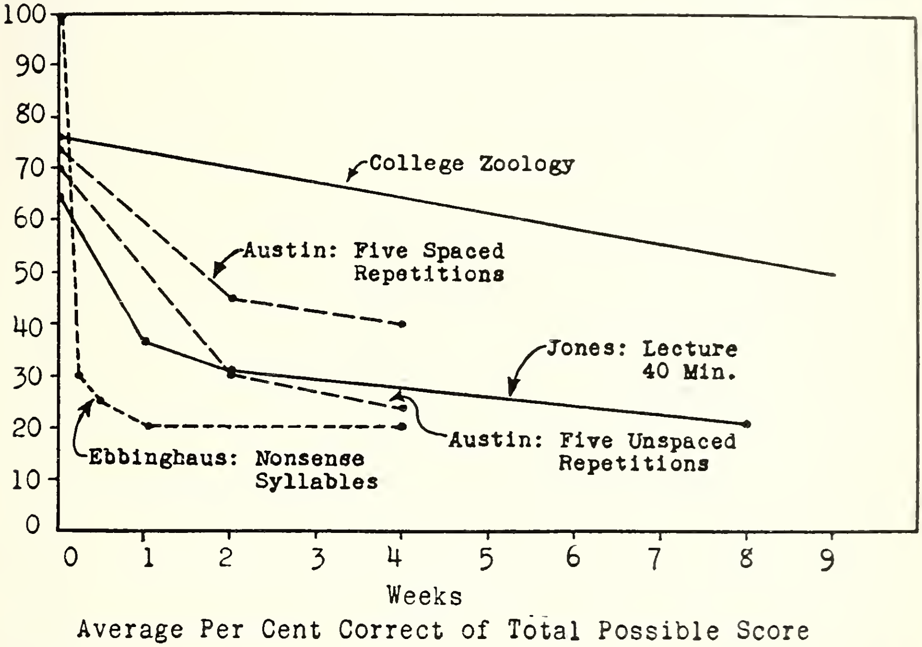


FIGURE 74. **Retention of Information in Adults: Comparison of Records Given by Jones, Austin, and Ebbinghaus**

by completion tests the retention of material presented in 40-minute lectures. His results show a rapid drop at first followed by a gradual loss until after 8 weeks only 24 per cent was retained. Greene (1931) measured retention of specific information from a college zoology course by sections of a final examination administered to different groups at 4-month intervals up to 20 months. A drop from 75 per cent to 50 per cent retention was recorded for the first 4 months, to 42 per cent for 8 months, to 37 per cent for 16 months, and to 32 per cent for 20 months. It was also reported that a mean of nearly 20 per cent of the items were known to students beginning this course. Similar findings have been reported by Johnson (1930) for college botany, by Greene (1931) for college psychology and chemistry, by Powers (1924) for high-school chemistry, and Noonan (1926) and Irmina (1928) for various elementary-school subjects. Nearly all of these reports show rapid forgetting at first followed by gradual decreases to approximately 20 per cent of the test items. Nonsense material, or items which involve complicated or rare meanings, are forgotten more quickly than the rest. None of these studies have carefully evaluated the mental factors involved in retention. Recall of specific verbal information seems to be the dominating activity, but in some instances methods of attending to or solving a problem and motor skills may be involved. A study by Swift (1910) reported nearly perfect retention of a motor skill – juggling two balls – after 6 years without practice. It is generally conceded that both simple motor habits and language habits, when thoroughly practiced, are retained for long periods.

These measures of amounts of retention are subject to all the errors that creep into appraisals of practice and growth. The per cents of items retained is a very rough measure which ignores the differences in quality and difficulty of the items in a test.

Another result of forgetting may be a change in the dispersion of scores in a group. Tilton (1936) reviewed 39 published studies most of which were reports of retention of verbal information from school studies. He found a marked tendency for the SD's of groups to increase during forgetting until the poorest members of the group ceased to forget any more. Thereafter the SD's decreased while other persons were still reducing their retention scores.

The writer believes that these results may be typical only of measures in which SD's had decreased in size with learning – a frequent finding in achievement tests which fail to appraise the more able students accurately. Where SD's increase with learning it seems probable that they would decrease with forgetting. It is possible, however, that in many instances retention will be longer among the better learners.

The measurement of individual curves of forgetting seems to be nearly unexplored except in a few laboratory studies. Doubtless, persons differ in rate of forgetting owing to a number of causes, such as variations in methods of study, amounts of overlearning, motivation, and distractions. Correlations between achievement scores taken at intervals of 3 or 4 months usually range from .40 to .60 for small groups of students. Jones (1922) found a mean coefficient of .57 ± .12 for equivalent tests of a lecture, when one test followed the lecture immediately and the other came 8 weeks later. Johnson (1930) found correlations of .75 and .84 between final examination scores and part of the same examination given 3 months later among small college groups. Greene (1931) reported correlations of .41 ± .12 and .43 ± .06 between two trials of the same test in psychology and chemistry, respectively, among small groups of students who had had the summer months in which to forget their instruction. These figures can do little more than whet the appetite for more accurate studies.

21

MEASUREMENT OF GROWTH AND SENESCENCE

I. THE NATURE OF GROWTH

Individuals begin life as a single cell, called a *zygote,* which has been created from a combination of two germ cells, one from each parent. The zygote divides very rapidly at first, thus producing many cells which are apparently similar to each other except for their positions. The cells on the outside of the animal soon develop a different structure from those on the inside. It is believed by biologists such as Morgan (1929) and Conklin (1922) that each cell of an individual has exactly the same inherited possibilities of development. The process of cells becoming different, called *differentiation,* must be due to the fact that cells develop in different positions and receive different kinds of stimuli. Differentiation of cells produces eventually several fairly independent systems, such as the circulatory, locomotor, receptor, nervous, and glandular systems. The differentiation of cells leads to the development of more specific patterns of behavior, called *individuation.* Individuation is illustrated by changes from gross surging responses of large muscle groups to specific responses of small muscle groups.

Patterns of individuation depend upon and often follow very closely the differentiation of cells. One classic illustration of this is shown in the work of Coghill (1926), who examined salamander embryos with a microscope and compared the growth of nervous connections with the behavior of the animals. He found that the behavior pattern always followed a particular cellular development. For instance, a nonmotile stage was followed by a flexure away from the point of stimulation only after particular nerve fibers had grown into position. He emphasized (1930) the fact that all individuation arises from the division of total reaction patterns into smaller partial patterns. Other illustrations of individuation are:

1. Gradual oral development of words out of poorly articulated sounds
2. The selection of a precise set of movements, such as those made in typewriting or in driving a car, from roughly co-ordinated movements, many of which are unnecessary
3. The formation of a precise meaning of a word from a vague meaning

Both differentiation and individuation are developed through *maturation* and *exercise* or *use*. Maturation is defined as development which takes place without exercise, but where the animal is nourished and kept in a healthy state. Exercise is defined as a series of experiences which lead to changes in behavior patterns through the retention of changes in structure. Maturation is often found to control to a large extent the variety and total number of cells in an organism, while exercises regulate the condition and size of the cells. The words *exercise, practice, learning,* and *use* are all used with similar meanings in this book.

Both differentiation of cells and individuation of responses are regulated in part by the body as a whole. This regulation for unified growth or action is called *integration*. It is accomplished at the early stages of development by physiological gradients in groups of cells. Later, integration of more distant parts of an organism is accomplished by chemical agents in the blood stream and by nervous impulses which initiate and accelerate or check various responses. These forms of regulation of the amounts and direction of energy expenditure are very complex. Growth must be thought of as the development of various structures and patterns of behavior in co-operation or competition with one another.

In testing one person it is usually impossible to separate the modifications of behavior which are due to maturation from those which are due to learning. Both types of changes seem to have similar curves of improvement, both are permanent only under certain conditions, both are probably concurrent with structural changes in the organism, and one probably affects the other. Some writers cannot distinguish any real differences between maturation and learning, but many investigators think that important differences may be described as follows:

1. Modifiability: Behavior or structure due to maturation is often more difficult to change than behavior due to learning. All other things being equal, this is measured by the number of repetitions necessary to produce a change in behavior.

2. Permanence: Behavior due to maturation is more lasting than that due to learning. It does not show rapid loss, as in some types of forgetting.

3. Regularity of Change: Behavior which is due primarily to maturation shows more regular growth changes. The results of learning, such as memory for specific events often show sudden and marked changes.

4. Time of Life: Improvements which take place in adults, after maturation has theoretically ceased and before senescence has a marked effect, are due to learning. Maturation is much more rapid in infancy than in later years, while learning seems to be nearly as rapid after adolescence as before.

5. Types of Processes: Changes of behavior, involving new reactions of a specific complex sort following a specific experience, are usually learned.

Of these criteria, this last one is probably the most commonly used. It is generally accepted that learning is present when an animal responds toward a goal in a more precise and efficient manner on a subsequent trial than on the first trial. Chapter 20 dealt with studies in which practice is apparently the major factor in improvement, and this chapter deals with studies of growth in which maturation is considered to be the principal factor in development.

II. CONDITIONS ESSENTIAL FOR GROWTH STUDIES

Before studying the typical results of measures of maturation, it will be well to consider conditions which will make such measures comparable and accurate. The following conditions are considered so essential as to be axiomatic:

1. The same processes must be measured throughout the growth curve (constant quality). This condition is basic to all growth measurement. Scores on a test become incomparable if they are determined by one type of response at one age and by another type at another age. This is a very common source of error. Thus, many and perhaps most of the form boards are difficult at the lower ages because the youngster has a hard time in perceiving similarities between the shape of the block and the shape of the hole into which it must be placed; however, at the higher ages the perception of these form relationships may be so easy that the test is principally a measure of speed of movement. Growth curves of stature are also often incomparable through changes in relative size of component parts of the body.
2. The same persons must be measured (constant native ability). In order to predict an individual's development we should have very clear evidence that his own individual curve is of a particular nature. In many measures of human structure and behavior it has been very much easier to get growth curves which show the averages of age

groups than to get consecutive measures of the same individuals. The result has been that we have a fairly large number of norms showing means and standard deviation of various age groups, but relatively few of individual growth curves over long periods of time.

Many investigators have pointed out that since averages may conceal the facts we cannot generalize about individual growth curves from average curves. It is possible, and rather probable in many cases, that the individual growth curves differ markedly in shape from the average growth curves of the same individuals. This situation is true when one person's acceleration at one period may be matched by another's retardation in such a way that the average shows no unevenness.

3. The units of measurement used must not vary in size or difficulty (equivalent steps on a measuring scale). This condition is necessary in order to avoid false conclusions regarding relative amounts of improvement. Many tests which are constructed with harder items at the advanced age levels do not give more credit for a hard item than for an easy item. Such tests indicate that various age groups are more similar in ability than they actually are.

This problem of scaling items presents a dilemma. On the one hand, we wish to avoid combining results; and on the other hand, it seems desirable in the case of measures of behavior to secure equivalent steps in a scale on the basis of group dispersions. The problem of scaling raw scores so that there will be equivalent steps is discussed in Chapter 22.

4. Environment must remain the same (constant degree of practice and motivation). This condition is also very hard to attain. Most tests are so much influenced by practice or special drill that an increase in score equivalent to several years' growth may be attained in a very short time (see Chapter 20).

5. True zero points and maxima must be known. This condition is essential for the accurate calculation of cycles of growth and percentages of individual development. The true zeros of mental development have not been measured directly, but Thurstone (1928) and others have devised methods of calculating zeros which have certain advantages. These are discussed in Chapter 22.

Measures which lack one or more of the essential conditions are to be looked upon as preliminary studies.

III. TYPICAL RESULTS OF GROWTH STUDIES

A consideration of the conditions necessary for measuring growth may lead one to despair of ever securing accurate appraisals. Rather such a consideration should lead one to caution in interpreting results too broadly, to an understanding of seemingly contradictory results, and to continued research. Some measures of growth have yielded standards for various populations which have considerable value when interpreted carefully. Typical measures of growth of anatomy, motor co-ordination, thinking, and other traits are given below to allow their comparison.

A. Anatomical Structures

Figure 75 shows average increases in weight for the whole body, for the brain, and for three endocrine glands. Before birth, all show similar curves, but after birth a remarkable divergence occurs. The brain and the thymus gland increase very rapidly, the total body weight more slowly, the testes remain nearly the same size for six years, and the adrenal glands rapidly lose weight. At about 12 years of age, the brain has reached its mature weight, the thymus is nearly twice the adult weight, and the other three, total weight, testes, and adrenal glands, are beginning to show the rapid increases of adolescence.

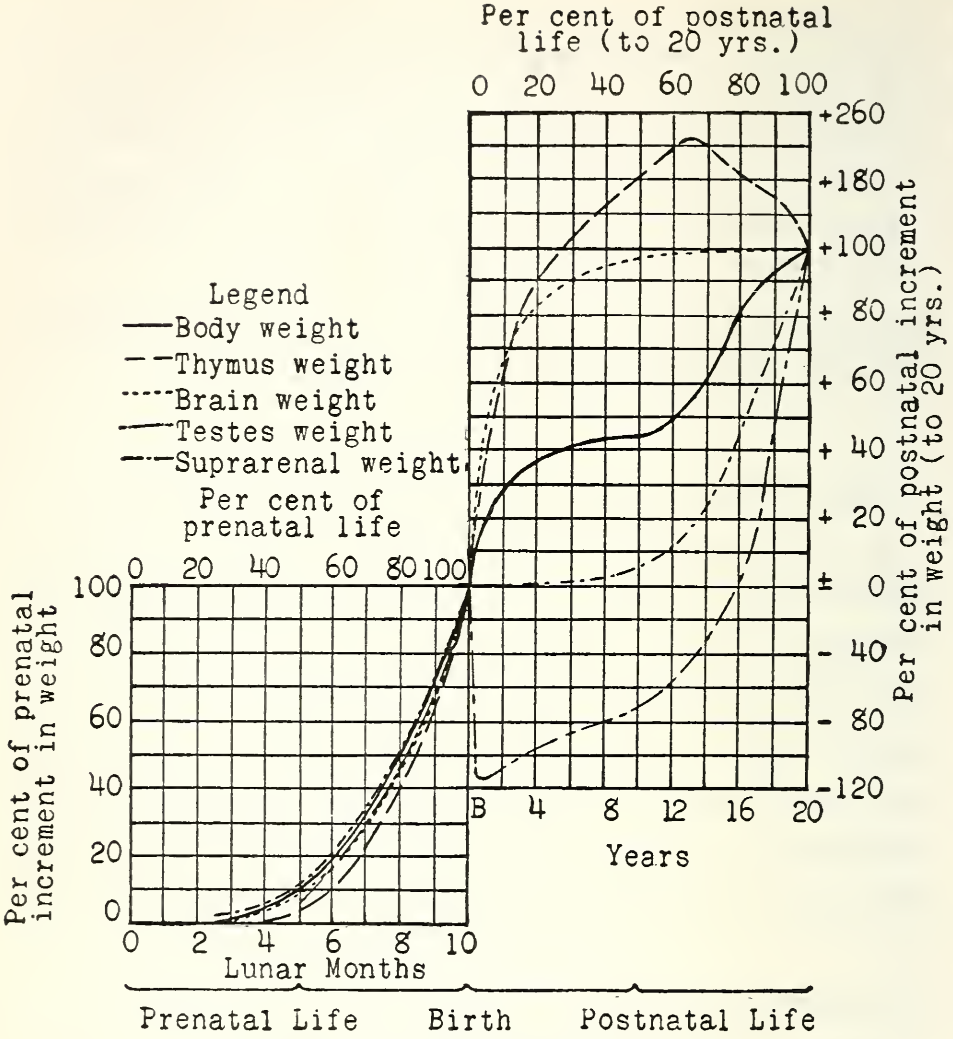


FIGURE 75. **Growth of Structures**

(From Harris *et al.*, 1930. By permission of the University of Minnesota Press.)

Figure 76 shows average curves for the calcification of wrist areas — a measure commonly recorded in studies of growth. Although all of the curves are fairly similar, some elements have reached half their maximum growth before others have started to ossify.

Figure 77 A shows the average curve for eruption of teeth. This is interesting since teeth can be observed rather easily and accurately, and since they are thought to be an important

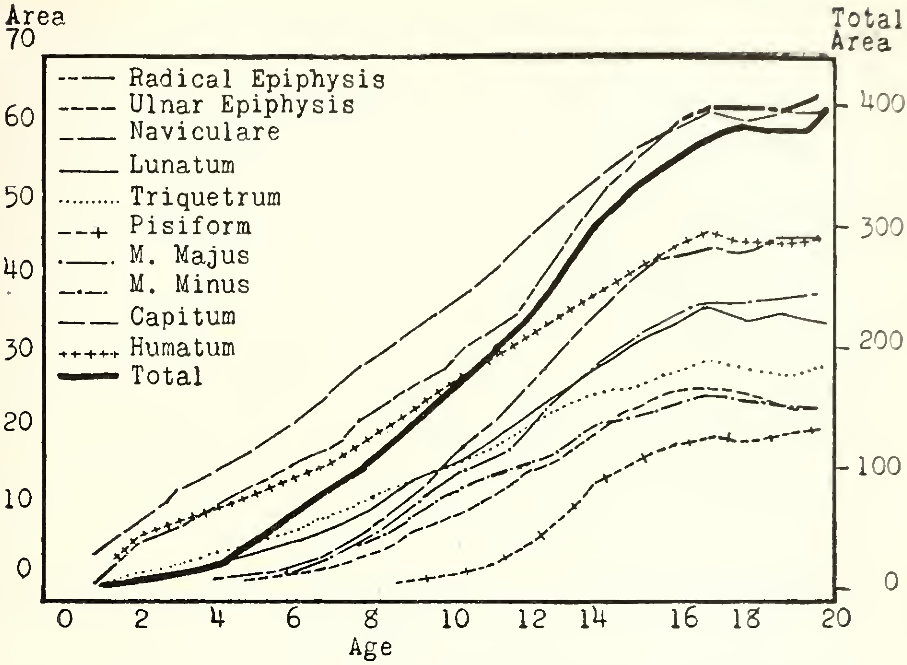


FIGURE 76. **Average Growth Curves for the Carpal Bones and Two Epiphyses for Laboratory School Boys**

(From Flory, 1936. By permission of the Editor, Society for Research in Child Development.)

index of physical maturity. The deciduous teeth reach a maximum at 4 years and begin to be discarded at about 5 years. The permanent teeth show three cycles: the childhood, the adolescent, and the adult.

Figure 78 shows records of height and weight of four groups of different intellectual abilities. Up to about 14 years of age the idiots lag behind the other three groups. Thereafter, changes in the curves appear, particularly in height. These changes are probably due to sampling errors among the moron and imbecile groups, for these groups were reduced by the discharge of many able-bodied and emotionally stable persons from the institution at about the age of 15 years. The idiot group was not so reduced, since its members require institutional care. The growth curve of the idiot group is very similar to the growth curves of normal

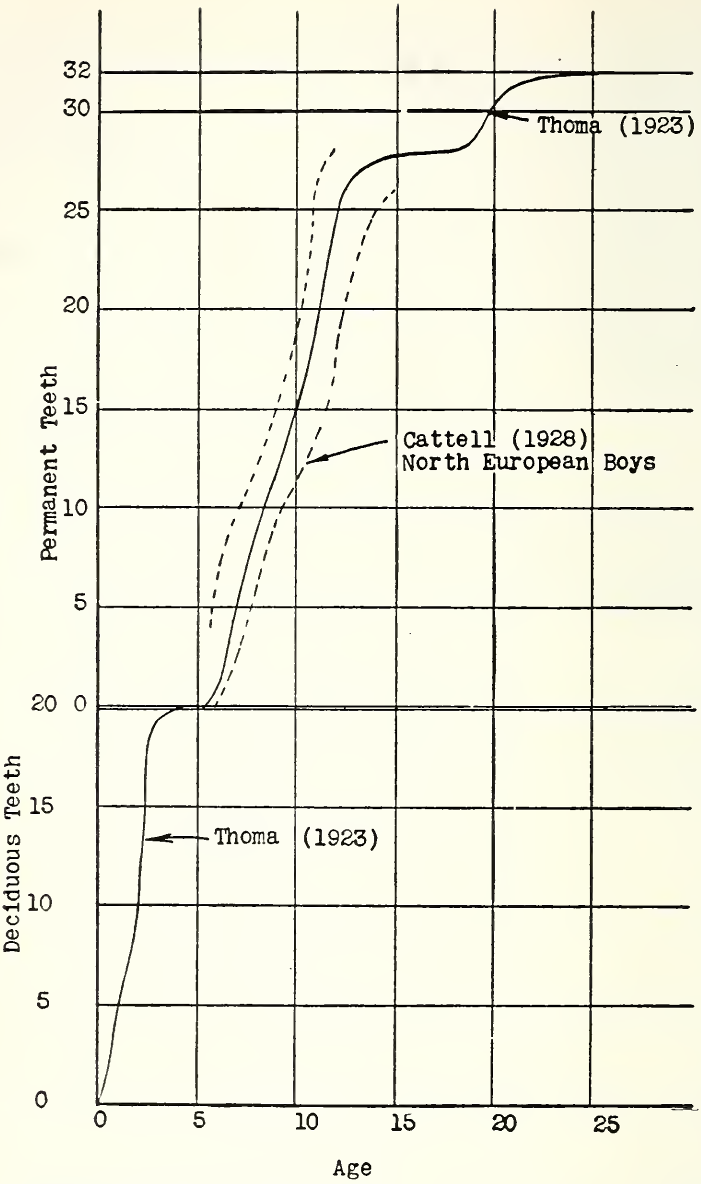


FIGURE 77A. **Dentition: Average Age of Eruption of Teeth**

children. This table was presented to show how sampling errors are introduced.

Figure 79 shows an individual height curve and the average curve for a group of Iowa boys. The average shows three very slight changes in the slope of the curve, the first

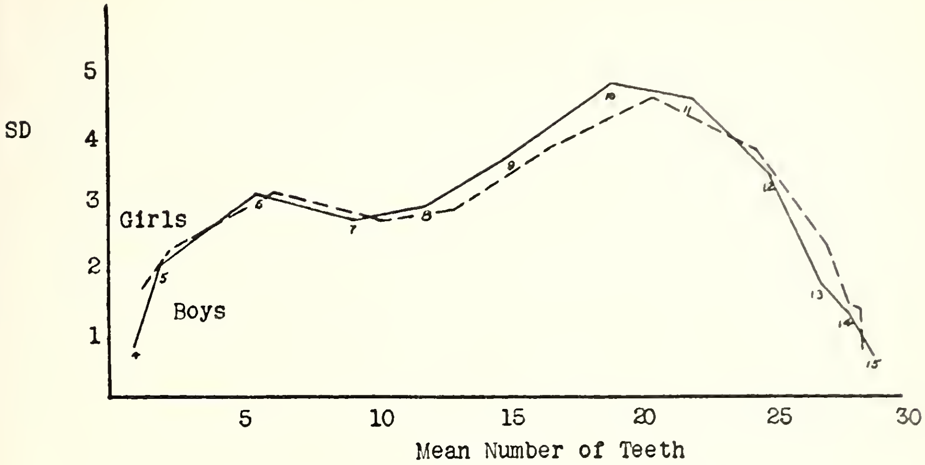


FIGURE 77B. **Permanent Teeth Cut, City B**

(After Courtis, 1937.)

of which is more marked than the other two. The individual curve shows much more marked inflections and suggests that the average curve may not be typical of any of the individuals who compose it.

B. Motor Co-ordination and Spatial Thinking

The results of Miles (1931) are usually taken to be representative of average growth and decline of motility and spatial thinking, although the numbers of persons were small, varying from 90 to 320 for each decade. Figure 80 gives some of his results in order of the amount of senescent loss. Curve A, which shows the least decline, shows the per cents of the maximum mean score for five age groups on total scores from several eduction tests. These tests required comparison and judgment of size and position when no time limits were imposed. Curve B shows the relative speed of dexterity and reaction time tests. Curve C indicates immediate memory span as measured by spatial learning and code tests. Curve D shows speed of visual span of perception.

Among these curves there is an inverse relationship between early maturation and early decline. The thinking

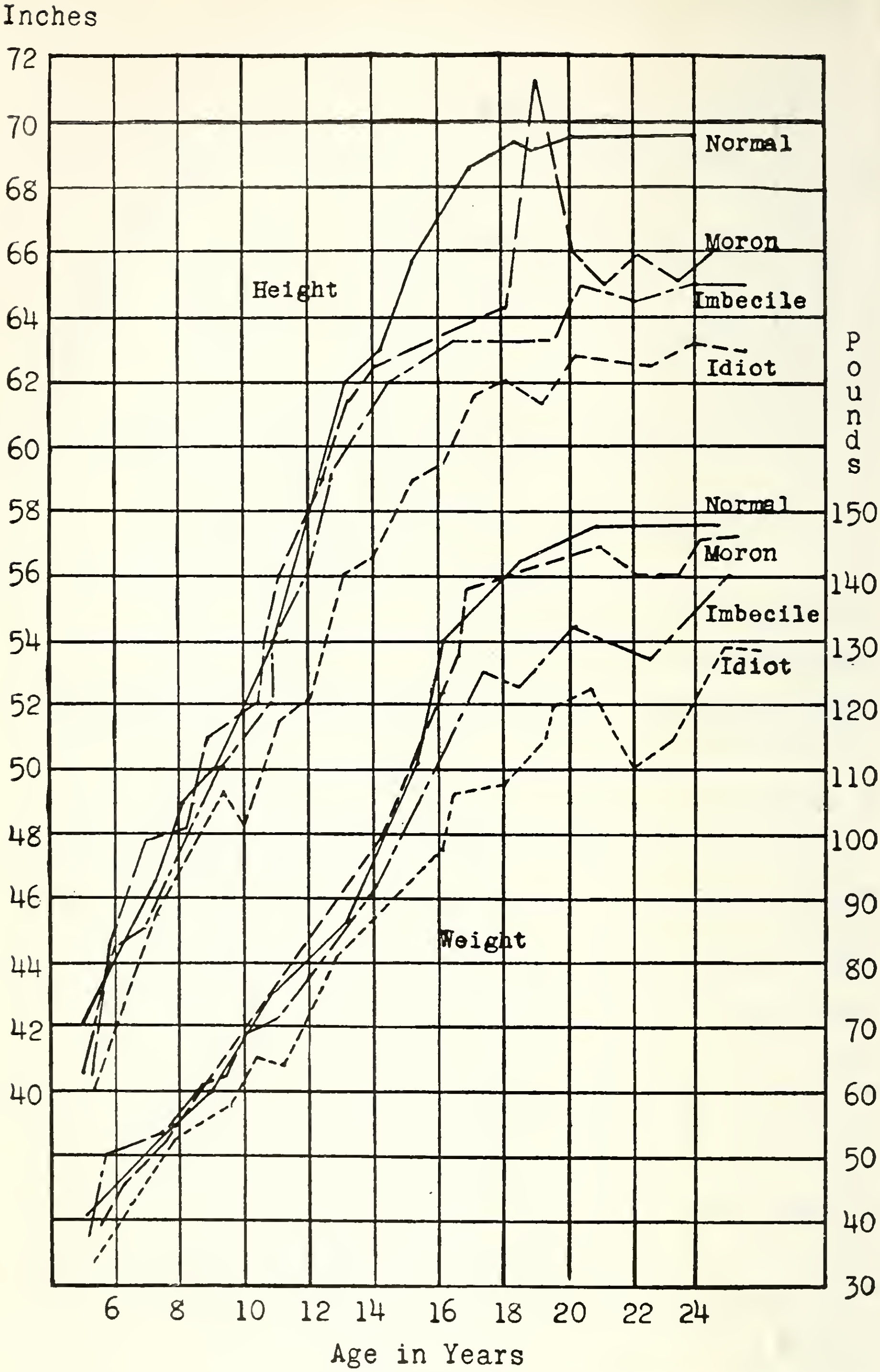


FIGURE 78. **Height and Weight of Feeble-Minded Males**

(Goddard, 1912)

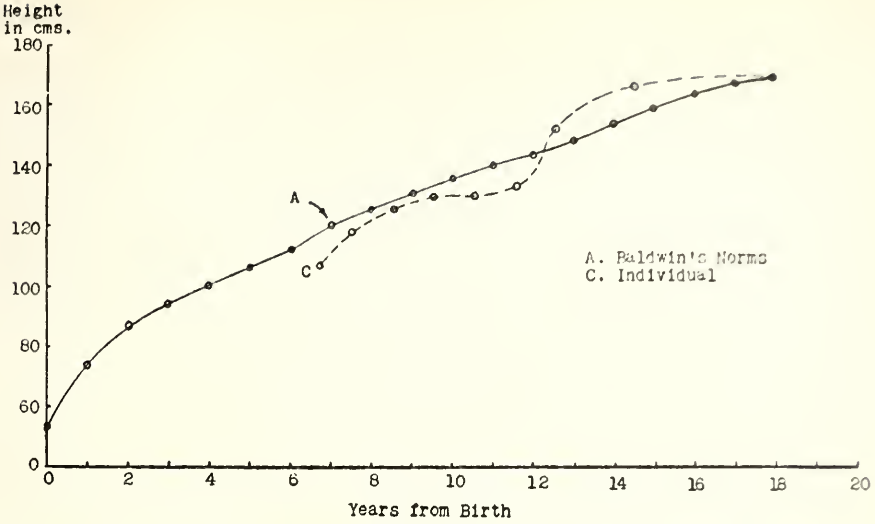


FIGURE 79. **Height of a Group and of an Individual**

(From Courtis, 1938.)

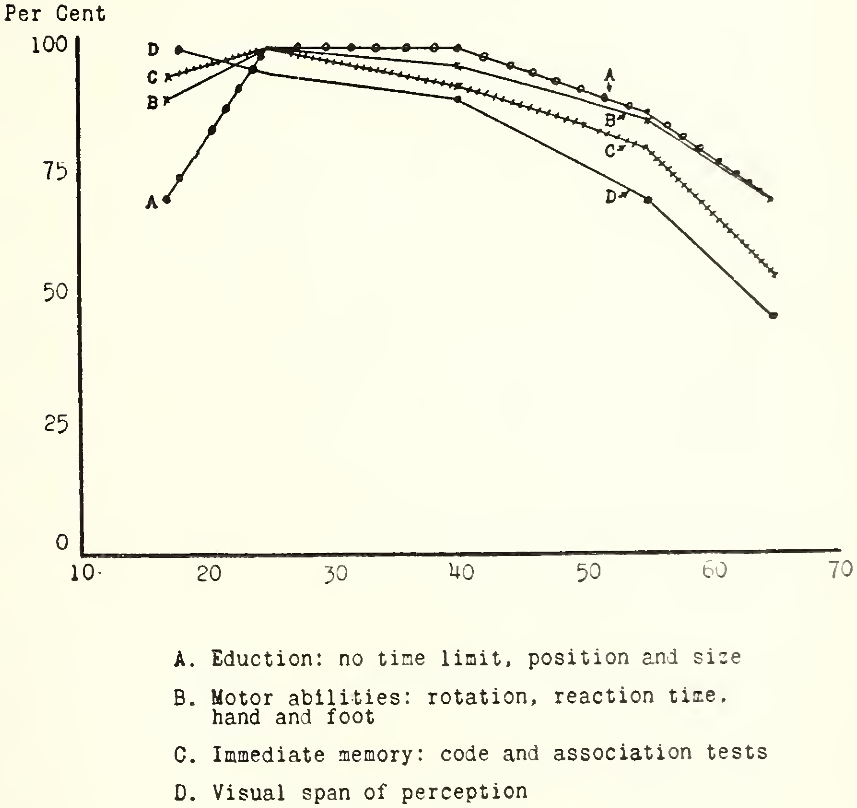


FIGURE 80. **Growth**

(Drawn from Miles, 1931 and 1933. By permission of the *American Journal of Psychology*.)

activities represented by Curve A mature later than the other activities and also decline later. The opposite is true of speed of perceptive ability, Curve D. These results led Miles to believe that tasks which were dominated by speed of perception or rapid movement suffered greater losses through age than tasks which required a careful and unhurried weighing of evidence.

C. Intellectual Growth

A study was made by Freeman and Flory (1937) of 269 pupils who entered the Chicago Elementary Laboratory School at about 8 years of age. The difficulty of securing continuous measures is demonstrated by the fact that only 3 took the tests for ten years, 19 for nine years, 47 for eight years, 75 for seven years, 66 for six years, and 58 for five years. The Stanford-Binet IQ's of this group had a mean of 115 and an SD of 13.61. Four types of tests were used throughout the study:

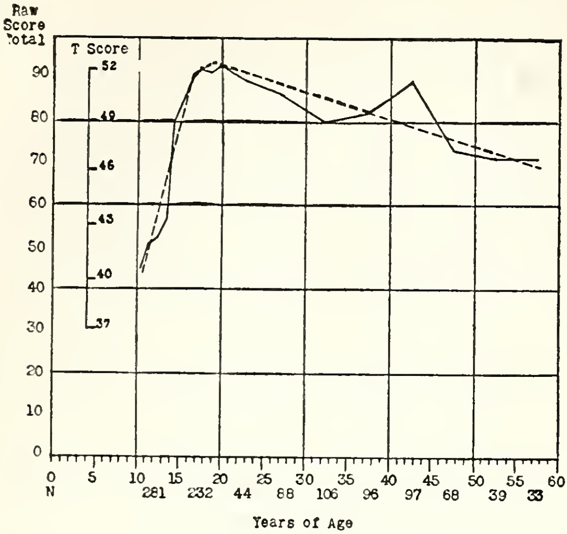
1. The Oral Vocabulary Test of 100 words from the 1916 Stanford-Binet Test.
2. Verbal Analogies Test: Three approximately equal forms were constructed. Each form contained 60 items, which were equated for difficulty according to the SD values furnished by an earlier standardization by Pintner and Renshaw (1920).
3. Verbal Completion Test: Three forms were made by selecting items which had been previously standardized by Kelley (1920) and Trabue (1916). Each form contained 21 items.
4. Verbal Opposites Test: Three forms were selected from material previously standardized by H. A. Greene (1918).

All of these forms were administered without time limits and no credit was given for faster work. The Vocabulary and Opposites Tests were given and answered orally. The Analogies and Completion Tests required written answers to a reading situation. All of the tests were given individually. The scores were the number of items correct according to keys, except that the completion test was given 2, 1, or 0 credits. The Vocabulary and Opposites Tests also allowed half-credits for imperfect answers. The results show that:

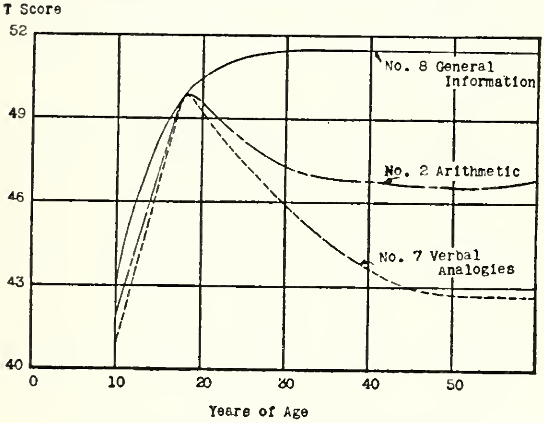
1. The annual increments of the middle and the highest third of the group decreased faster between the ages of 9 and 16 than did the annual increments of the lowest third at the same chronological ages. This disparity means that the highest pupils had scores a little nearer the lowest as they all progressed in age. This fact was particularly noticeable among the boys, but not among the girls. The investigators found that this tendency for the brighter pupils to mature earlier than the others is clearer in the Analogies and Opposites Tests, less clear in the Completion Tests, and not at all clear in the Vocabulary Test.
2. There was a very slight superiority of girls at the ages of 11 to 13, and of boys at the ages of 15 to 17 years.
3. Correlations between retests one year apart given at 11 and 12 years were as follows: Vocabulary, .826; Analogies, .639; Completion, .658; Opposites, .633; total score, .845; Stanford-Binet, .64; and Otis, .79.
4. At the age of 12 years, the Vocabulary correlated with the Opposites Test, .72; with Analogies, .48; and with the Completion, .52 for 166 boys. Similar correlations were found for the girls. Obviously there is a marked similarity between Vocabulary and Opposites. All of these tests correlated nearly .56 with the Stanford-Binet.
5. Correlations between tests given five to ten years apart were so low that good prediction of later ability could not be made from the earlier tests. Relatively large individual differences in rates of growth were often reported over periods of five or more years.
6. The age of maximum development could not be determined from this study, for increases in scores were found up to the highest age studied. The form of curve seems to indicate that the maximum would be reached somewhat after the age of 19 years.
7. The effects of learning could not be distinguished from those of maturation in this study.

A report by Jones and Conrad (1933) is typical of nearly all the reports on mean intelligence test scores. and it also illustrates some of the results of averaging growth curves of several different tests. Figure 81 A shows the mean total

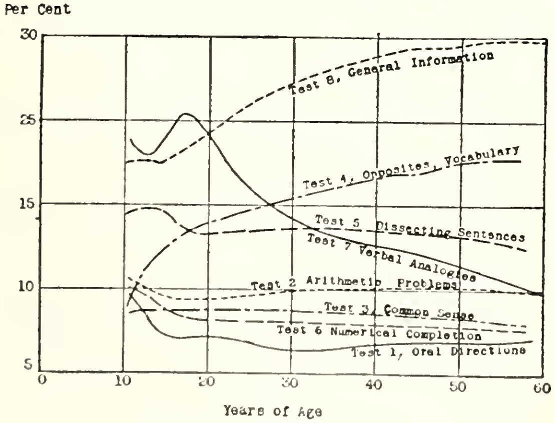
United States Army Alpha scores for various age groups and the number of persons in each group. The actual curve shows marked irregularities which are probably due to inadequate sampling of various age groups. These irregularities have been eliminated by smoothing, a procedure which might introduce slight errors in the shape of the curve. Figure 81 B shows the average growth curves for three of the eight tests. The other five tests fell between these. All



A. Amounts Correct on Total Test



B. Amounts Correct on Subtest



C. Per Cent of Total Score

FIGURE 81

Growth and Decline of United States Army Alpha Test Scores 1191 Children and Adults in Rural New England

— Actual ······ Smoothed

(Arranged from Jones and Conrad, 1933. By permission of the Editor, *General Psychology Monographs*.)

three have similar curves from 10 to 18 years, but they differ enormously after age 18. The Verbal Analogies test, which involves speed of easy inferences to a large degree, drops down to the 10-year-level, or approximately .7 of the SD of the 18-year-group. The Arithmetic Calculations test drops .3 SD and the General Information test rises approximately .2 SD before the age of 30 years. One might be tempted to conclude that the total curve was equally affected by the three tests from 10 to 18 years, but the relative contribution of each test to the total score cannot be known from a diagram which shows only a few of the tests. Figure 81 C shows the per cents of the total score contributed by each test at each age. Verbal Opposites, Information, and Arithmetic tests contribute increasingly larger amounts up to the 18th year. At 60 years they account for 56 per cent of the total score, whereas at 10 years their total was only 37 per cent. The other tests contribute correspondingly smaller amounts with advancing age. These figures indicate the presence of several independent factors and show very convincingly that the total scores are not similarly composed from age to age. Jones and Conrad point out that one cannot expect to secure valid analyses of growth from total scores of such data, but that scores must represent similar skills for each year.

A report of Miles (1932) is diagrammed in Figure 82 to show that both average and superior IQ groups seem to lose ability as measured by a 15-minute Otis Self-Administering Test. In the test, speed of recall and verbal comparisons are important. The curves show a gradual loss of nearly 20 per cent in IQ from the ninth to the seventieth year. These results point to the desirability of using a standard score or centile ranking instead of an IQ for adult groups.

Three curves are presented here which have to do with reading skills. Figure 83 and Table 10 show the mean scores on a ten-minute speed-of-reading test for Grades III to XVI. Figure 83 also shows the spread of the scores by lines indicating 1, 2, and 3 SD's above or below the average. The

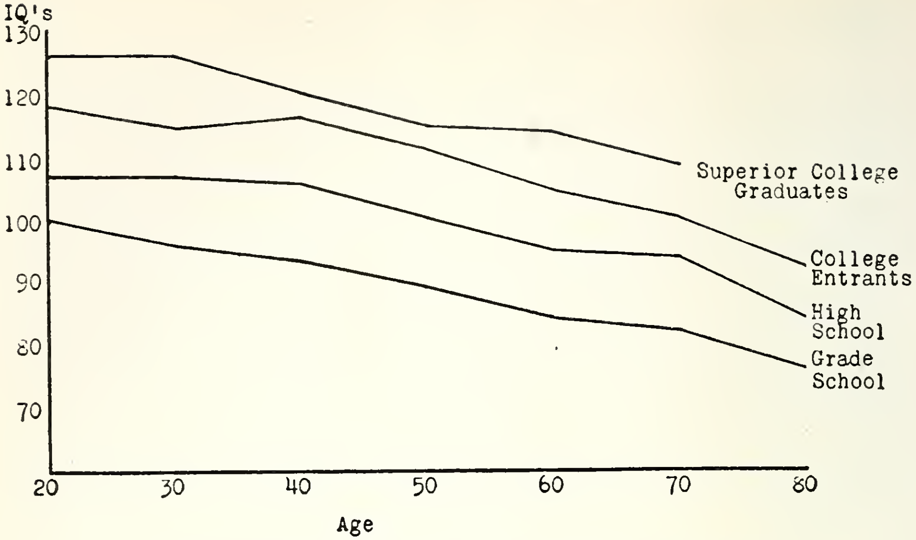


FIGURE 82. **Decrease in IQ's with Age**

Otis 15-Minute Test

(Miles, 1932. By permission of the *American Journal of Psychology*.)

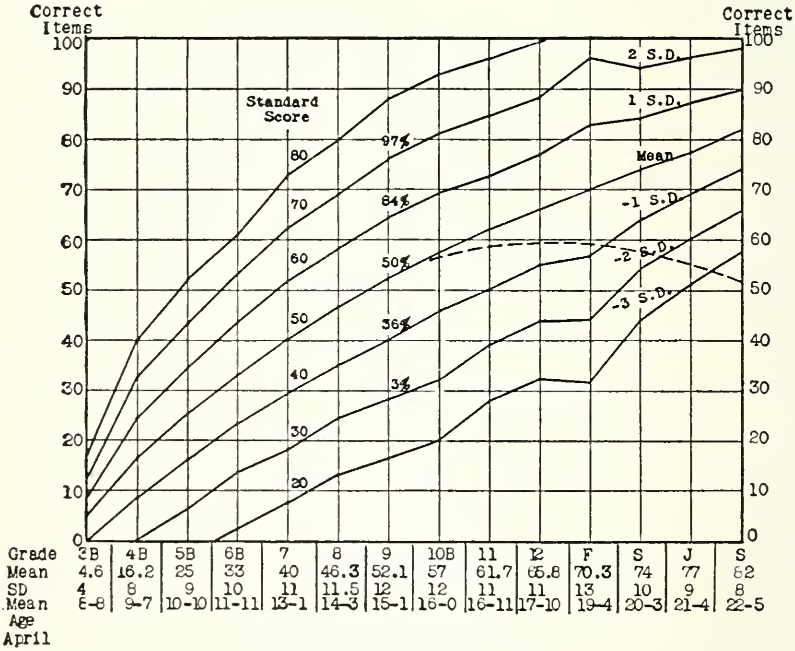


FIGURE 83. **Grade Norms, Michigan Speed of Reading Tests, Form I or II**

Correct Items, 10 Minutes

(Greene, 1934)

lines present a fairly symmetrical dispersion which increases from the third to the ninth grade, decreases slightly to the twelfth grade, increases for college freshmen, and decreases sharply during a college course. The increases in standard deviations through the ninth grade are generally considered to be due principally to maturation. From the ninth to the twelfth grade, the decreases in SD's are probably due to the elimination from high school of the poorer readers. This elimination also raises the mean scores in these grades. The increase in SD's from the twelfth to thirteenth grade is probably the result of sampling college freshmen who come from many different schools, some of which stress speed-of-reading, while others do not. The decrease in SD's through college may occur from a sampling process which eliminates the slower readers. Since all throughout the chart the effects of training are doubtless in evidence, one is not justified in assuming that training has been constant. The average adult score was found to be approximately 54, or equal to the 12-year-level. This chart illustrates very clearly the difficulties of securing comparable samples of a population over a period of years.

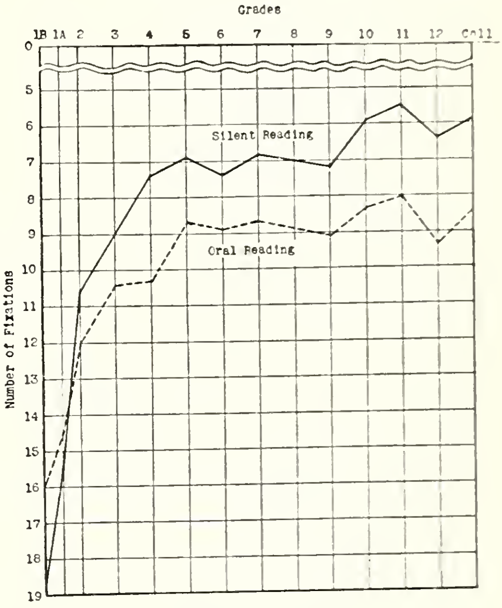


FIGURE 84

Growth Curve of Eye Fixations

(Redrawn from Buswell. 1922. By permission of the University of Chicago Press.)

Figure 84 shows that the mean number of eye fixations used in reading a standard passage was at first greater for silent than for oral reading. Near the end of the first grade the silent reading fixations became fewer and the difference is marked thereafter. Figure 85 shows three individual growth curves in reading, which indicate the presence of marked cycles in this function. Individual A matures earlier than B or C and also reaches a higher maximum than C. Individual B has a longer second cycle than C and reaches a higher maximum. Figure 86 shows two MA curves for individuals. Both of them show inflection points near the eighth year, whereas an average MA curve does not show such inflection points. These curves again raise the question, "How typical of individual ability are mean curves for a group of individuals?"

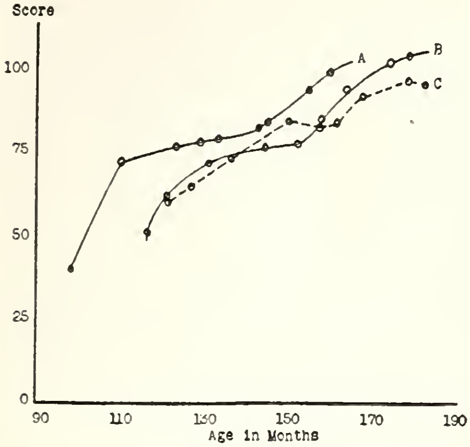


FIGURE 85

Individual Growth Curves in Reading

(From Courtis, 1938.)

Studies of growth in early childhood are illustrated by

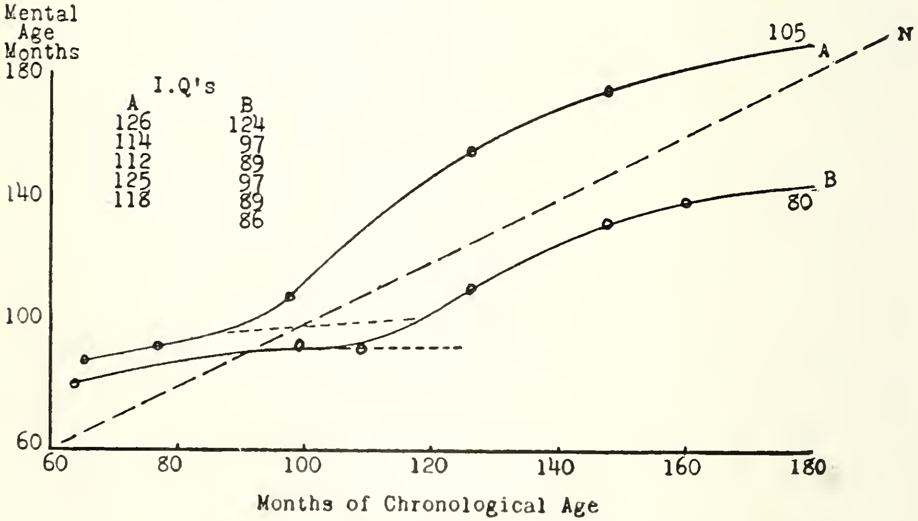


FIGURE 86. Mental Age Curves

(Courtis, 1938.)

mean points on the California Scale in Figure 87. The raw scores and scores on an absolute scale are compared. Similar figures from Stutsman (1931) are available for the Merrill-Palmer Scale. The interpretation of these is difficult since the composition of the scales at various ages is not known.

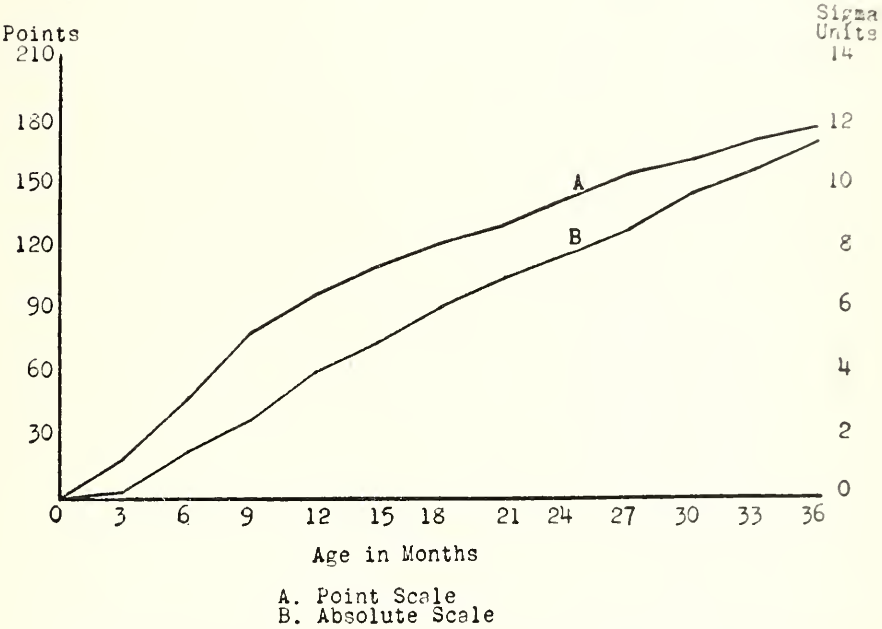
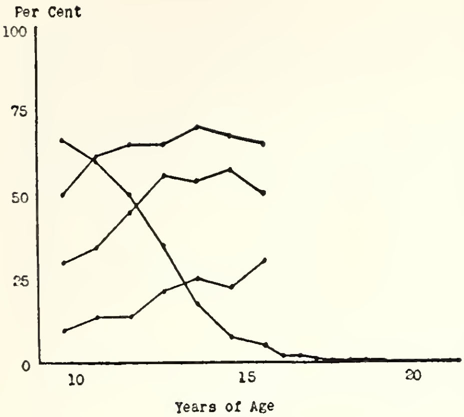


FIGURE 87. **Growth in Infancy and Early Childhood**

(Bayley, 1933. By permission of the *Genetic Psychology Monographs.*)

The curves rise with a slight negative acceleration which is reduced by absolute scaling to almost a straight line.

Another approach to the appraisal of mental growth is the recording of per cents of persons in various age groups who pass an item. Such records are given in Table 25 and Figure 27 for items in the Gesell Scale, in Table 53 and Figure 38 for Binet-test items, and in Figure 34 for the spelling of single words. Records of this sort are used for the absolute scaling of items. They show the growth of an item in a population, rather than the growth of a person in the achievement of items like the previous curves. Both sorts of curves report aspects of the same phenomena.



Fi (Percentage of Age groups who reported the activity in a given week.)
activity in a given week)

FIGURE 88

Changes of Interest with Age

(After Lehman and Witty, 1927. By permission of A. S. Barnes Co.)

D. Character Traits

A fairly large number of studies have been made of the incidence of interests, attitudes, and character traits. Typical of these is Figure 88, which shows changes of interest with age in city and rural children. Three of the items show rising curves, but one, playing with dolls by girls, has a dropping curve. Another interesting report is that on nail-biting in Figure 89. The curve rises from zero at 3 years to 20 per cent of both boys and girls at 4 years, and later to 45 per cent of girls at 12 years, of boys at 14 years.

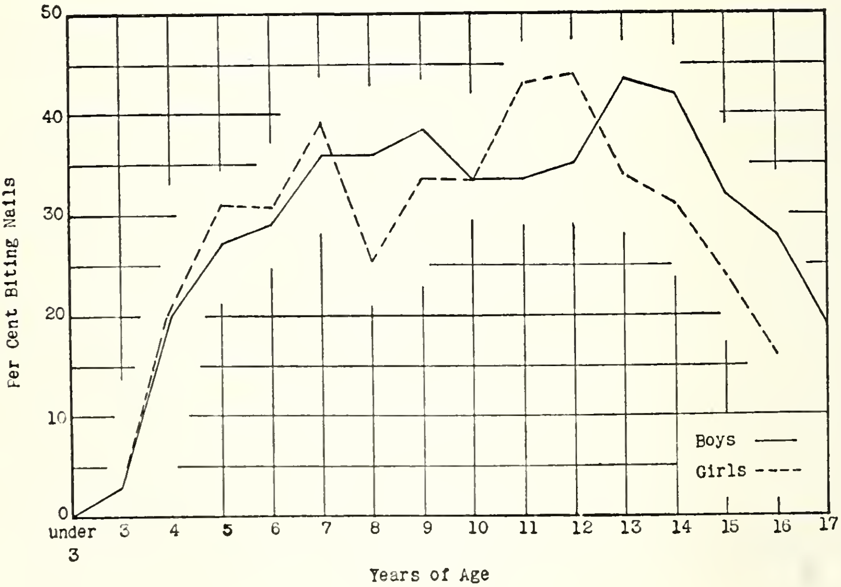


FIGURE 89. **Incidence of Fingernail Biting**

(Wechsler, 1931. By permission of the *Psychoanalytic Review*.)

Thereafter a rapid drop occurs. Girls reach 15 per cent at 16 years, and boys 20 per cent at 17 years.

A third report deals with the results of tests of masculinity and femininity, Figure 90. There appears to be a marked increase in masculinity among boys from the eighth to the eleventh grade. Thereafter the average male score becomes less masculine until at 60 years he is nearly at neutral on this scale. From the eighth to the thirteenth grade the girls' scores show a rapid change away from femininity toward neutral, but thereafter they move slowly toward femininity, changing very little after 40 years of age.

Similar results, shown in Figures 60 and 66, give further evidence that a number of trait patterns wax and wane independently of one another.

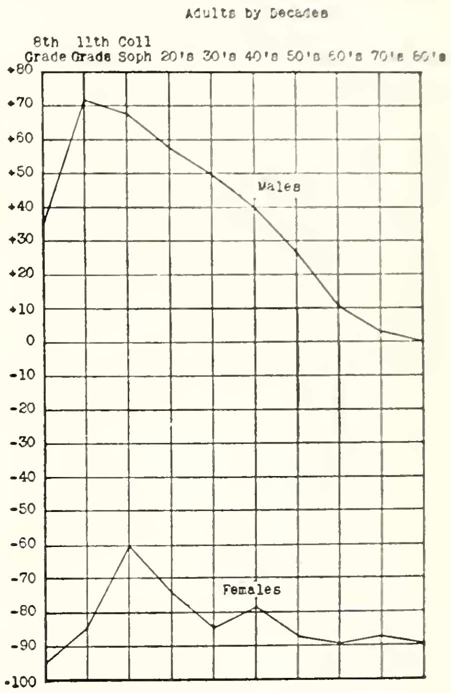


FIGURE 90

Masculinity-Femininity Scores

(From Terman and Miles, 1936. By permission of the McGraw-Hill Co.)

IV. DISPERSION OF SCORES

Group studies of growth often lead to the question of whether distributions become more or less widely spread out with age. Figure 83 illustrates a common tendency for SD's to increase with age, and also a tendency to vary with sampling. Figure 91 is more free from sampling variations, as nearly the same persons were retested for nine years. It shows that SD's of weight and height increased from 8 to about 14 years, and then decreased slightly. The decrease

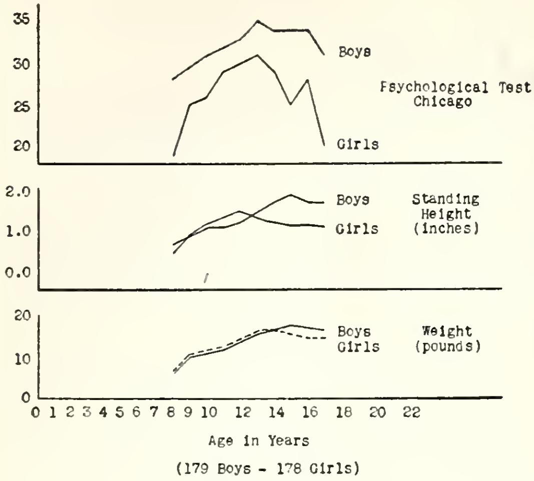


FIGURE 91A

Standard Deviations and Age

(Redrawn from Abernethy, 1936. By permission of the Society for Research in Child Development.)

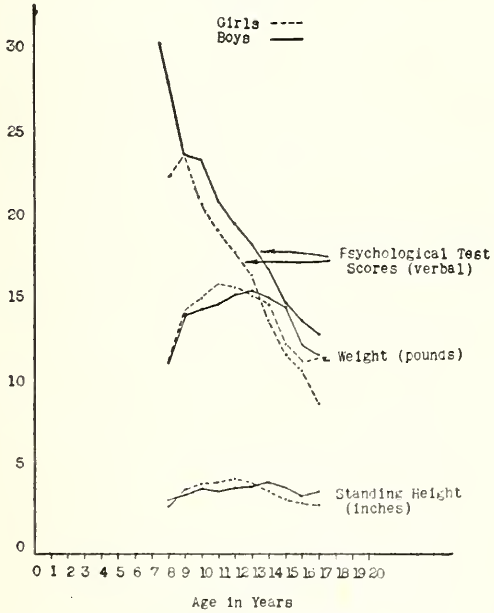


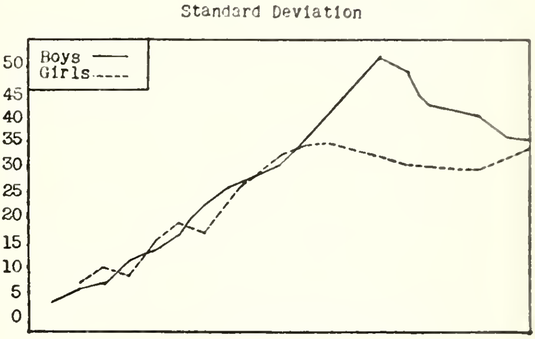
FIGURE 91B

Coefficients of Variation

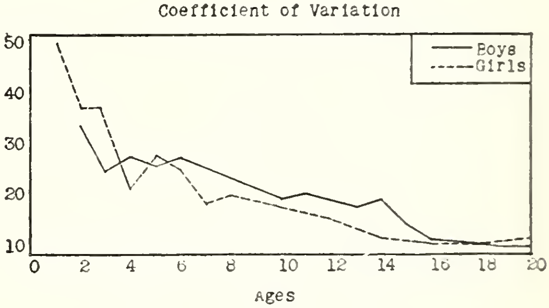
(Redrawn from Abernethy, 1936, Figure 7. By permission of the Society for Research in Child Development.)

began earlier for girls than for boys. The SD's of scores on the Chicago Psychological Test increased markedly from 8 to 13 years, then dropped back almost to the 8-year-level by the age of 17 years. This drop in SD's is probably not typical of tests of this sort and it cannot be readily explained in this case. It might be due to a selection of testing items which were either too easy or too difficult at the advanced ages. Both of these situations would result in a bunching of scores. Or it might be a reflection of changes in motivation and special training of a portion of the group. The boys showed greater variation than the girls — a typical finding.

In order to compare the *relative dispersion* of two variables, another measure of dispersion is frequently sought which will show how much of the total mean score is included in one SD. The SD is divided by the mean and the resulting per cent is called the *Coefficient of Variation* $\left(CV = \frac{\sigma\ 100}{M}\right)$. Figure 91 B shows the CV's for the data given in Figure 91 A. The measures of heights have nearly the same CV's for all ages from 8 to 17 years. The measures of weight have CV's nearly three times as large as those for height. The CV's for weight rise from 11 per cent at 8 years to above 15 per cent from 10 to 15 years, and then drop back to 12 per cent at 17 years. The CV's for the psychological tests are very high at first, but show a fairly regular drop throughout the period from 8 to 17 years so that the later ages have CV's of about 13 per cent.



A. Variability of Total Area of Carpal Development



B. Variability of Carpal Development

FIGURE 92

Osseous Development in the Hand as an Index of Skeletal Development

(Redrawn from Flory. 1936. By permission of the Society for Research in Child Development.)

Figure 92 diagrams the SD's and the CV's for carpal ossification. The SD's have fairly steady increases to age 11 among girls, and 14 among boys. The CV's decrease regularly with age from 3 years to adulthood. The adult value, which is approximately 10 per cent lies between the CV's for height and weight of adults.

Figure 77B shows two marked cycles of SD's for permanent teeth. The first cycle runs from 4 to 8 years and the second from 8 to 15 years. This figure illustrates very nicely the effect upon SD's of a scale of measurement in which all persons are growing toward approximately the same maximum. In this case the maximum is the number of teeth. Barring accidents, all adults should have 32 teeth; hence the SD of the group is zero.

After this inspection of relative dispersions, the need for more research is clear. The marked fluctuations in CV's shown in most of the charts should be further investigated to find out whether the changes are due to genuine growth phenomena or to the techniques of measurement which are likely to be present. Three causes of variations are often reported:

1. An increase in CV's may occur when a group changes from a slow period of growth to a fast period, if too wide age groupings are used as the unit. Thus, if the rate of growth from 8 to 9 years is slow, and that from 12 to 13 is rapid, a group of persons whose ages range from 8 to 9 years will have a smaller CV than the same group when the ages extend from 12 to 13 years. If, however, the CV's were calculated for age groups of only 1 month, then the CV's would not show much variation owing to grouping of persons of different ages. This possible explanation of increases in the CV's of height and weight during the adolescent period needs to be confirmed by research.

2. A marked decrease in CV's is generally observed, as in the psychological test data, Figure 91 B, when the raw-score zeros are considerably above the developmental zeros. In this scale the raw-score zero would be typical of 5- or 6-year-old performance, but the language and thinking skills needed in the test were probably developing rapidly from birth. Chapter 22 describes one technique for calculating a developmental zero which yields CV's for mental test scores that are nearly constant for wide age ranges.

3. Variations in CV's from age to age may also be caused by the roughness of the measure, as in the growth of dentition, Figure 77 B. The total number of teeth is such a rough measure that all persons in a group have the same score at certain ages, when the teeth certainly differ widely in size, shape, and texture.

V. RELATIONSHIPS AMONG VARIOUS GROWTH PATTERNS

Two lines of investigation are open for discovering the relationships among various growth patterns. One involves correlations of test scores and the other compares growth curves for various abilities of an individual.

A. Correlational Studies

The correlations reported in Chapters 9, 10, and 11 lead to the conclusion that, in large random samples of any age group, measures of structures, abilities, and traits are only slightly related. Further evidence is summarized by Garrett (1938) indicating that test scores of verbal, number, spatial, and immediate memory skills are less highly correlated in adults than in children. Brooks (1937), in reviewing studies of growth from 2 to 12 years, concluded that there was overwhelming evidence that mental growth rates of nearly average individuals tend to differ considerably from the average, but very superior or inferior children tend to remain in their classifications. On the other hand, Lincoln (1936) found that changes in IQ's were greatest for the children with the highest IQ's over a ten-year period, as shown in the Harvard Growth Study. The results of repeating the same battery of tests after an interval of three years, using 79 boys and 82 girls, were reported by Asch (1936), Tables 50 and 51 (see page 255).

These studies neither demonstrate nor exclude the existence of marked relationships among growth patterns, but one is made to believe that other techniques of discovery may be more fruitful. There are two procedures which seem likely to show how growth patterns are related in individuals. One of these comes from anthropology. It is suggested that various homogeneous racial groups be studied separately, since patterns of growth in one group may differ considerably from those in other groups. Possibly, high relationships that will appear in a homogeneous group will be covered up by averaging several racial groups. Since it is hard to define a race, the selection of a single racial group is a perplexing problem. The results of careful mental measurement may contribute to more accurate discrimination. The other suggestion is that individual curves be studied, so that the concurrence and order of pattern development may be noted. Further studies that follow these suggestions will undoubtedly contribute a clearer picture of correlations among growth patterns.

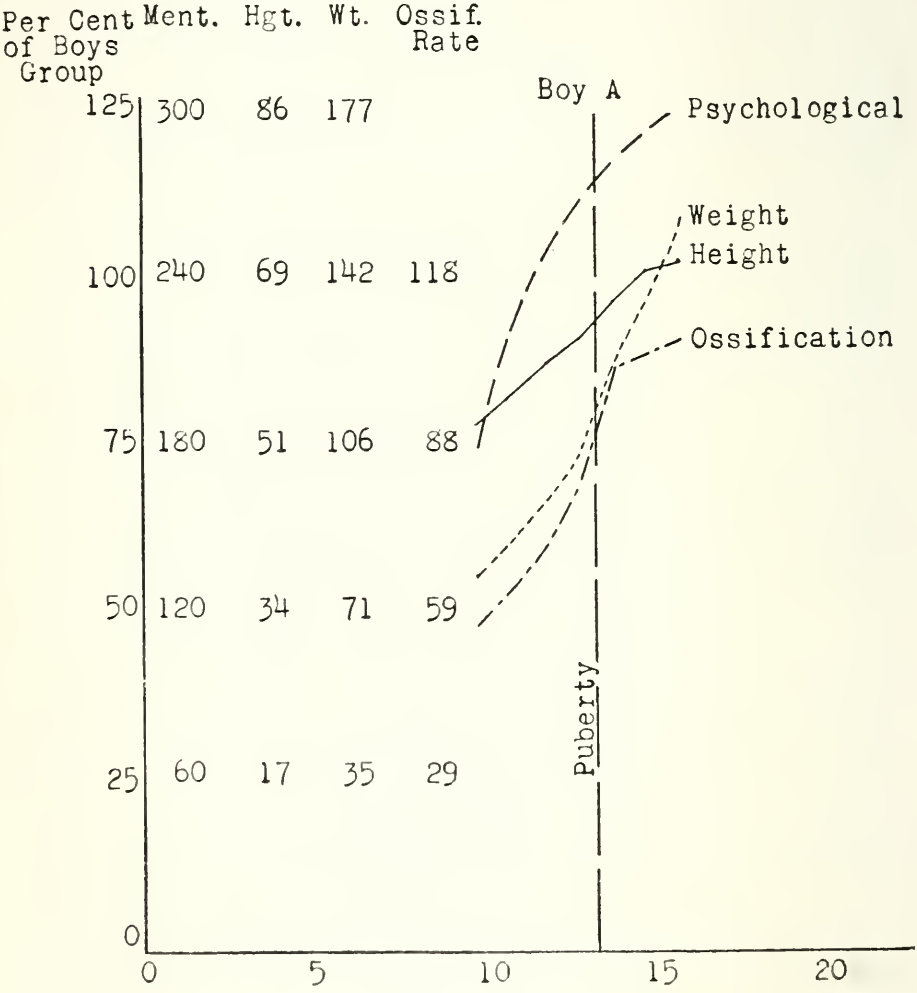


FIGURE 93. **Individual Age Increments**

(Redrawn from Abernethy, 1936. By permission of the Society for Research in Child Development.)

B. Individual Growth Curves

Three figures illustrate the comparison of growth curves in the same individual. Figure 93 shows a case from the Chicago Growth Study which illustrates raw score differences in growth curves. The height of this boy developed in nearly a straight line, the psychological scores were negatively accelerated, and the weight and the ossification measures were positively accelerated to the 15th year. There was no adolescent acceleration in mental ability corresponding to that in weight.

Figure 94 shows three charts of individual growth in which age norms are used. Each boy's scores are changed into age equivalents and charted. Boy A is above average on every measure except carpal ossification and is most advanced in mental age. Boy B becomes more and more retarded at later ages in every measure except height. Boy C shows a marked difference between rates of development of reading and of dental age, but most of his patterns are fairly constant. Olson (1937) has found that children with patterns like

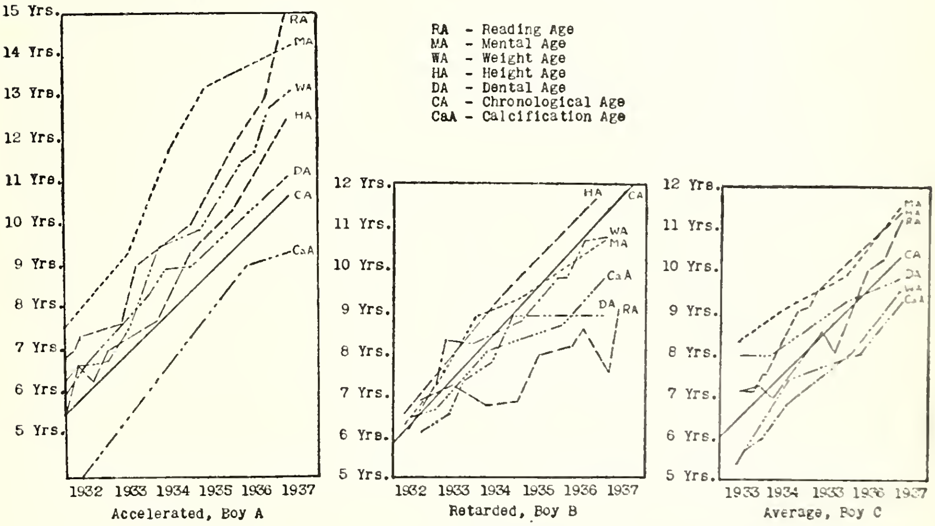


FIGURE 94. **Three Individual Growth Charts**

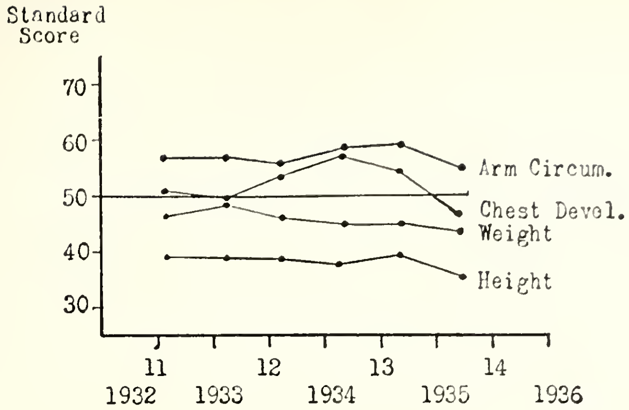
(By permission of W. C. Olson, 1937.)

Boy C usually have a larger number of behavior difficulties than the normal child. These difficulties seem to be due to the wide scattering of abilities.

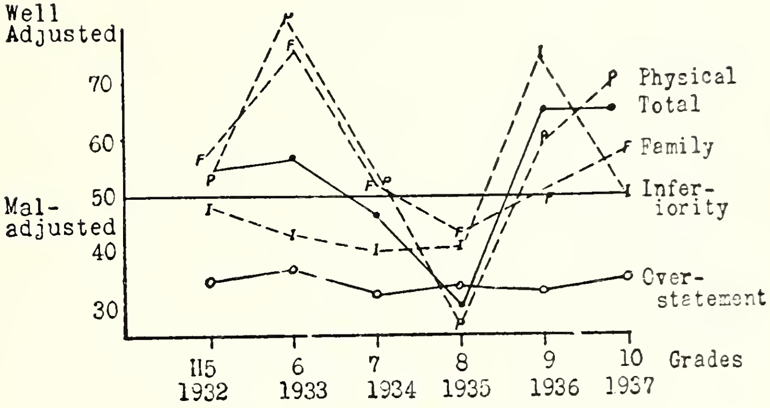
Figure 95 shows repeated appraisals of one boy by means of standard score charts. The variables shown were chosen from a much larger number on file at the Institute of Child Welfare of the University of California. The physical measures indicate that the boy was small for his age but nearly normal in weight and slightly above normal in muscular development. The emotional inventory, which was very thorough, shows a period of maladjustment at about 13½ years. The ratings of adjustment show that he was always untidy, childish, and very active, and that he gradually lost his good looks, popularity, and older friends. His fighting activities were at first far above average, then moderate, and later, again excessive. Together, these charts picture a boy who was able to compete favorably in Grades V, VI, and VII, but then lost his equanimity through a combination of home conditions, physical smallness, and inability to continue in some of the academic courses.

Studies of this sort are fascinating fields for research. Among the important problems which await solution are the following:

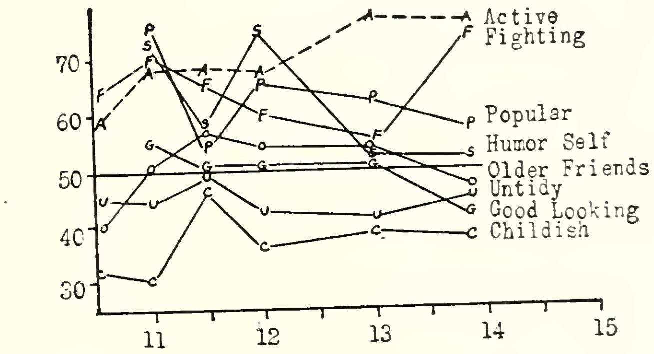
1. Do any of these curves show regular cycles which are not shown by the group averages?
2. Do any functions predict changes in other functions a year or so later?
3. Is there evidence for compensation of deficiencies by growth in other directions?
4. Are some measures good indicators of a number of other variables?
5. Are converging and diverging curves of growth related to periods of maladjustment?
6. Which method of scaling gives the most significant comparisons?



A. Physical



B. Emotional Inventory



C. Adjustment and Environment

FIGURE 95. **Individual Growth Curves**

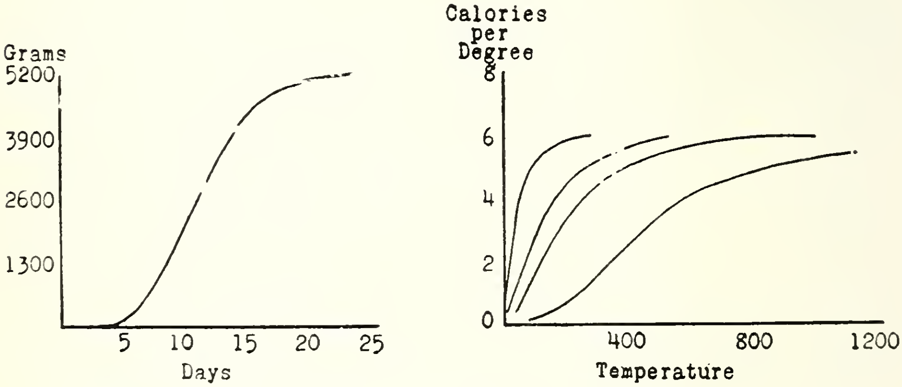
(Drawn from data furnished by H. E. Jones from the files of the University of California Child Welfare Research Station.)

VI. THEORETICAL CONSIDERATIONS

Many of the growth curves which have just been described are somewhat similar in slope, but most of them differ enough from one another to suggest that their essential similarity may be doubted. Consequently, the question arises, "Is there a universal growth curve?" Courtis (1929) believes that there is, and that it can be found in nearly every sort of data when environment and native ability are held constant.

A. Simple or G Curves

Illustrations of very well-controlled situations are seen in Figure 96 A and B, which show the weight of a pumpkin and the atomic thermal capacities of various metals. In these experiments, variations in temperature, light, humidity, and electricity were reduced to approximately zero, or changed systematically. The results are a set of very regular curves which have the same shape. Curves are said to have the same shape when they can be expressed by the same formula, ignoring raw score differences. The reasons why all these curves are the same are not clear. The weight of a pumpkin



A. Weight of Pumpkin B. Atomic Thermal Capacities

FIGURE 96. **Simple Growth Curves**

(A. By permission of Courtis, 1932; B. Randall, Williams, and Colby, 1929. By permission of Harper and Bros.)

increases through an increase in both the number of cells and the size of the cells. A conversion of food results in new chemical combinations of elements. The total pattern is partly determined by the chemical substances in the germ plasm of the pumpkin seed. On the other hand, the atomic thermal capacities are indications of the number of calories needed to raise a particular mass of metal one degree. There are no increases in cells or chemical elements, but a rearrangement of particles, or movements of particles within the atom. A general principle which will account for the similarities in curves has been put forward by Courtis (1933). He states that all growth and practice must be conceived as a change of populations (or cells or atoms) in response to forces. If the boundaries and the forces remain constant, the elements in the population respond at different times depending upon their position in the group and their various resistances to change.

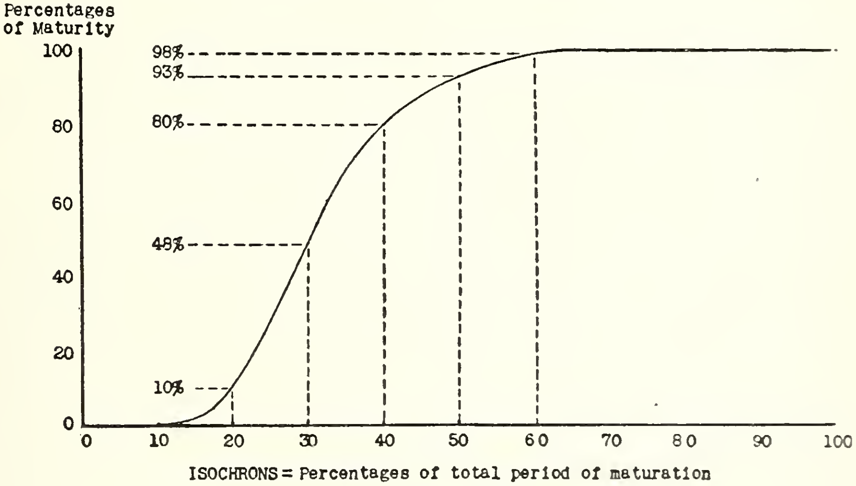


FIGURE 97. **G Curve**

(After Courtis, 1929. By permission of *School and Society*.)

A particular curve has become rather widely known as the *G* curve, shown in Figure 97. The same curve was found earlier by Gompertz (1825) in studies of growth of human populations in restricted areas. It has since been reported by

Courtis (1932) to apply to a large number of social and biological phenomena, such as:

1. The number of fruit flies that, beginning with two flies and ending after a definite time limit, exist in a small bottle with adequate air, moisture, and food
2. The number of men in each age group within a stable population who attain the married state
3. The elimination of students from college, by classes
4. The number of persons in each age group able to spell a given word or to pass a particular test item
5. The waxing and the waning of interests in age groups of boys
6. The number of boys in each age who can clear 4 feet in the high jump
7. The number of boys in each age who have a particular tooth
8. The number of automobiles of a particular model in use each year after manufacture
9. The effect of fraternity membership on scholarship
10. The effect of race upon college entrance
11. Many kinds of practice and learning of homogeneous materials under constant conditions

In view of the variety of the data sampled, it is a remarkable fact that all of the curves in the list above are practically identical in shape when the environment is held constant. Courtis suggests a definite law of growth, namely, given a certain germ plasm or structure, constant environmental forces will produce a regular change toward a definite maximum. The law, and the *G* curve, can be expressed by the formula, $Y = Ki^{rt}$, in which Y is the measured attainment at time t, K is the maximum development, i is the amount of development at the beginning or t, and r is the rate of growth or slope of the curve. When three points are known on a curve, K, r, and i may be calculated. The merits of this particular formula cannot here be compared with those of formulas proposed by Pearl and Reed (1920), by Thurstone (1919), by Robertson (1913), and by a number of others. Much research is needed, using carefully selected populations and good controls to check the rough measures of development now available. It seems reasonable to expect that a definite generalized curve of growth will be found, since essential similarities of biological phenomena have been widely observed.

B. Cycles: Successive Combinations

In actual measurement of persons or animals we seldom find complete, smooth curves such as those in Figure 96 A and B. Practically all curves of growth for anatomical or behavioral data show irregularities. If the *G* curve is typical of growth of a constant organism in a constant environment, then irregular curves may be due to the combination of several simple curves. Courtis (1938) believes this to be the best explanation, and he illustrates it with an analysis of maturation of height (see Figure 75). The first curve extends from conception to birth, the second from birth to the 7th year, the third from the 4th to the 14th year, and the fourth from the 10th to the 18th year. The adjacent curves overlap so that the end of one tends to cover the beginning of the next. The result is a fairly straight line for a considerable portion of the period of growth. Another curve, Figure 86, shows similar cycles for a girl repeatedly measured with the Stanford-Binet Scale. Courtis has found considerable evidence that height and mental ability develop in similar cycles, as shown in Figure 98, in which Mary Doe's growth curves are given for these two attributes.

It is quite possible that more extended studies will reveal individual cycles with fairly definite relationships between physical and mental growth. This finding is also to be expected from casual observations. Many children are known to grow 2 or 3 inches taller in a particular year and to develop a spurt in intellectual activities at earlier or later periods. Several research centers having been gathering individual records for several years, comprehensive reports on growth cycles should appear very soon.

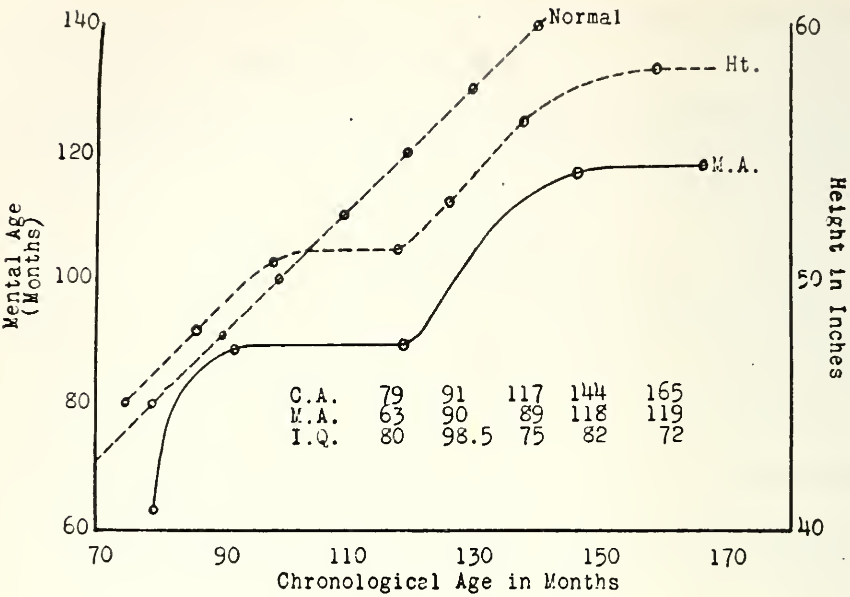


FIGURE 98. **Growth Curves for Mary Doe**
(Courtis, 1938)

C. Simultaneous Combinations

Another sort of combination of growth curves is shown by the mean curve in Figure 76. This exhibits the growth of the ossified sections of wrist-bone area shown on an X-ray photograph. None of the component curves very closely resembles the *G* curve or any other simple curve. Neither does the total area of carpal ossification secured by totalling these areas show a simple *G* curve. There are a number of possible reasons for these results. It may be that the scale of areas is not a good one to use, because wrist bones actually have three dimensions. Measures of volume might give different results. Or, since the wrist bones develop in conjunction with the forearm and the hand, it may be that total arm patterns are necessary in order to get an adequate picture of bone growth.

D. Suggested Scales

Mean curves of growth for intellectual behavior have generally shown fairly regular development in contrast to the

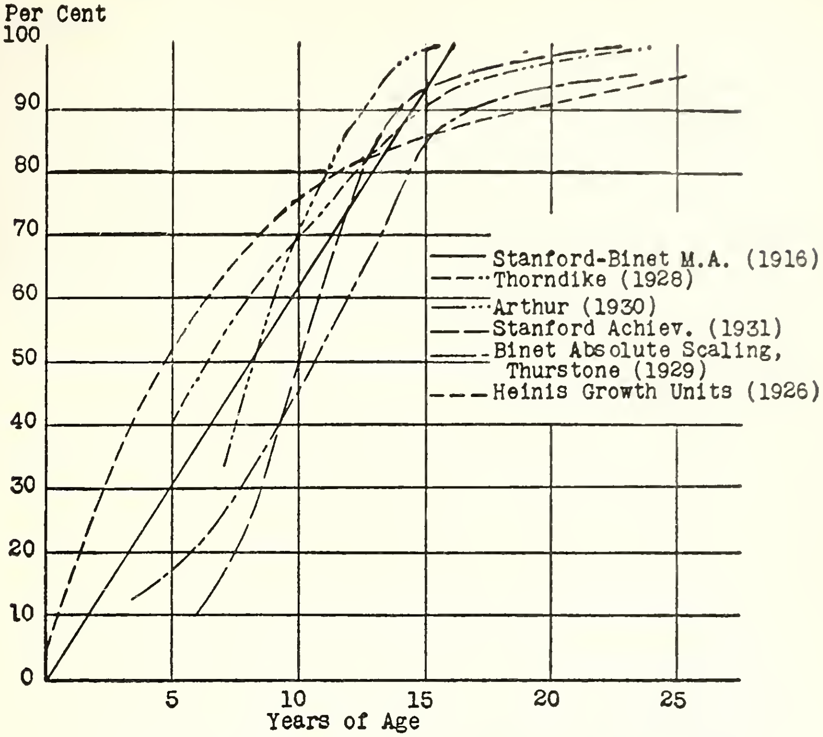


FIGURE 99. **Smoothed Mental Age Curves**

Per Cents of Maximum Average Achievement

curves for anatomical data (see Figures 75 and 81). This fact has led a number of investigators to suggest standard growth curves by smoothing their recorded results. Figure 99 shows several curves for different tests or for different populations. The mental age curve is a straight line by definition. The Stanford-Achievement Test curve is a theoretical scaling of school achievement tests, which is nearly the same in shape as Arthur's (1930) point scale but more retarded, owing perhaps to the early maximum of the Arthur scale. For the most part, Thorndike's (1928) curve is constructed from the CAVD tests. These were scaled in SD units. At first, the Heinis curve goes up faster than the rest, but after 15 years of age, more slowly. It has been extrapolated at both ends.

The differences among these curves are probably too great to be due to sampling errors alone. The Heinis curve shows the average individual to be 50 per cent mature at 4½ years, the Thorndike at 6½, the Stanford-Binet at 7½, the Arthur at 8½, the Stanford Achievement and the Thurstone Scaling of the Binet at 10 years. Space will not allow a critical discussion of the material. One effect of using absolute scaling for growth curves, shown in Chapter 22, would be to reduce the differences noted here.

22

STANDARD DEVIATION OR ABSOLUTE SCALING

I. THE NEED

One of the most difficult and persistent problems of mental measurement is that of securing scales which have equivalent units or steps and comparable zeros. Raw-score scales seldom have these attributes since they are made by adding together points which represent different qualities and quantities of skills. Chapters 2 and 5 discuss ways of avoiding qualitative differences in scale items. This chapter considers a method of making the points in a scale quantitatively equivalent.

Differences in the difficulty of items are often observable. For instance, in an altitude test such as the United States Army Alpha the items are purposely selected to increase in difficulty as one progresses through the test. Unmeasured differences in the difficulty of items may give rise to two sorts of errors:

1. An inequality of steps in the raw-score scale.
2. A failure to have all desired levels of difficulty adequately represented. A number of items may be bunched at one level of difficulty when none are found at another level.

Such errors may produce scales which fail to show the true relation among persons, and hence create distortion or skewing of frequency curves. It is, of course, highly desirable to avoid large distortions.

Two methods have been widely used to eliminate raw score distortions. One, called mental-age scaling, assumes that for any individual equal periods of growth in a constant environment can be used to secure equivalent steps in a scale. In this sense, the growth of one month at the age of 3 is considered to be equivalent to the growth of one month at 10 years of age. Mental-age scaling is discussed in Chapter 11. The other method, called standard deviation (or SD) scaling, is described in this chapter. SD scaling, which Thurstone (1925) has referred to as absolute scaling, provides equivalent steps in the sense that competent judges report the observed differences in behavior represented by scores of 1 and 2, as often as they report the differences between scores of 2 and 3, or 24 and 25, or a difference of one in any part of the scale.

Both methods of scaling have their limitations. Both are dependent upon a particular group of persons with particular training; hence they do not usually provide scales which can be used for other dissimilar groups. Both are rendered less precise by chance factors and ambiguities. Mental-age scaling is limited by the fact that mental increases due to growth are so small that they are very difficult to measure after early adolescence. Indeed, the higher levels of mental-age scales do not have true MA units. They usually have SD scale values which have been arbitrarily converted into MA's (see Chapter 11).

In constructing an altitude test, SD scale values of items can be used to select items to represent different levels of difficulty. This procedure is particularly valuable when two or more equivalent forms of a test are desired. Following Thorndike's tutelage, Ayres (1915) employed this method in devising a scale for spelling and Woody (1916), for arithmetic problems. Thorndike's (1926) CAVD Tests were selected according to SD scale values. In 1925 Thurstone published a more complete scaling technique which was followed by Richardson and Stokes (1931) in preparing an intelligence test and by Greene (1938) in preparing adult technical vocabularies.

II. THE NATURE OF SD SCALING

SD scaling of items is very much like the process of assigning standard scores to persons, described in Chapter 3. In the same way that a person can be given a standard score which represents his place in a group of persons, a test item may be given a standard score which represents its place in a group of items. A person's position is determined from the number of items he passes compared to the successes of the rest of the group. The position of an item is determined by the number of persons who pass the item, compared to the number of persons passing other items. Thus a person's skill is measured in terms of item difficulties, and item difficulty is measured in terms of successes of persons.

This method of evaluating the first variable from the second, and then reversing the process and evaluating the second from the first, may seem illogical if one expects to find somewhere in the universe a standard which is independent of anything else. The fact is, however, that this method is the one most frequently used by scientists. Forces or objects are always defined by reference to something else. Physical force is measured in terms of the mass which it can or does move, and mass is measured in terms of the force needed to move it. So also in measures of human behavior accurate scales can only be based upon operations in which one type of energy is changed into some other type by amounts which are judged to be equal to an arbitrary standard by a group of competent judges.

III. CALCULATION OF SD SCALE VALUES

The calculation of SD scale values for items is simple enough for one group of persons. The SD values for some of the Binet items used by Burt are shown in Table 112. The procedure given by Thurstone (1925) was as follows for the 7-year-group:

1. Administer all items to a large group selected at random.
2. Find the per cent of persons who passed each item, Table 112, Column 1, and Figure 100.

   Change the per cents to standard deviations, Table 112, Column 2. These are SD scale values. To avoid negative values, arbitrarily assign zero to the easiest item, and then place the others on the scale according to their distances from this item, Column 7.

When one wishes to have a continuous SD scale for items which are passed by all or none of one age group, but are passed by portions of other age groups, complications arise. The general plan is to use the SD of one group as a unit of measurement for all groups. This plan is carried out by weighting the SD's of all groups so that they conform to the accepted unit. First, the relative sizes of all SD's are found, and then each is multiplied by a number which makes it equivalent to the standard and hence to all the others. The procedure is as follows:

1. Secure SD scale values for each item for each group as shown in Table 112, Columns 3 and 4.

2. Find out whether the two sets of scale values have a linear relationship. If the groups have had similar training, the scale values of items from one group should correlate very highly with those from the other group. The two sets of scale values from Table 112 are plotted in Figure 101 to show their high correlation (.99). In this and similar cases Thurstone (1925) has shown that both groups may be represented on the same SD scale.

3. Find the means and sigmas of the scale values in the usual manner, Table 112, Columns 5 and 6, for each age group. Since these are based upon the same items, they orient each group to

TABLE 112

Absolute Scale Values for Burt's Binet Data

		1	2	3	4	5	6	7
No.	*Item*	*Per Cent Correct*		*SD Values*		*SD Values Squared*		*SD Scale*
		7 yr.	8 yr.	7 yr.	8 yr.	7 yr.	8 yr.	7 yr.
13	Draw Square	98.2	99.2	− 2.10	− 2.41	4.41	5.81	.00
21	13 pennies	96.6	99.0	− 1.83	− 2.33	3.35	5.43	.27
31	Right, Left	83.8	90.1	− .99	− 1.29	.98	1.66	1.01
34	Differences	70.6	80.0	− .54	− .84	.29	.71	1.56
38	Count 20 to 1	53.6	76.0	− .09	− .71	.01	.50	2.01
42	Months	35.3	61.6	.38	− .29	.14	.08	2.48
46	5 Weights	21.0	36.1	.81	.36	.66	.13	2.91
48	Draw Memory	9.8	28.9	1.29	.56	1.36	.31	3.39
55	Mixed Sent	0.9	15.3	2.37	1.02	5.62	1.04	4.47
	Sum *			− 14.64	− 40.27	113.55	119.32	
	Number of items			50	50	50	50	
	M.			− .29	− .80			
	S.					1.48	1.32	

* These figures resulted from summing 50 items instead of the nine items actually shown.

(Thurstone, 1925, Table 1. By permission of the Editor, *Journal of Educational Psychology*.)

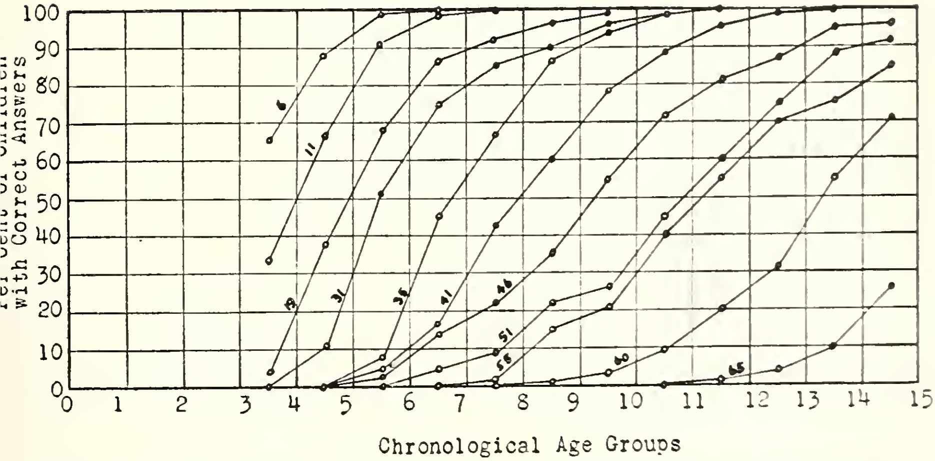


FIGURE 100. Per Cent of Age Groups Passing Each Binet Item

(Thurstone, 1925. By permission of the Editor, *Journal of Educational Psychology*.)

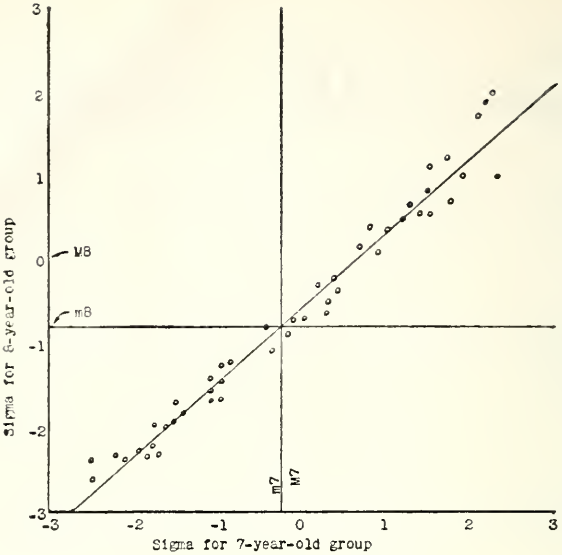
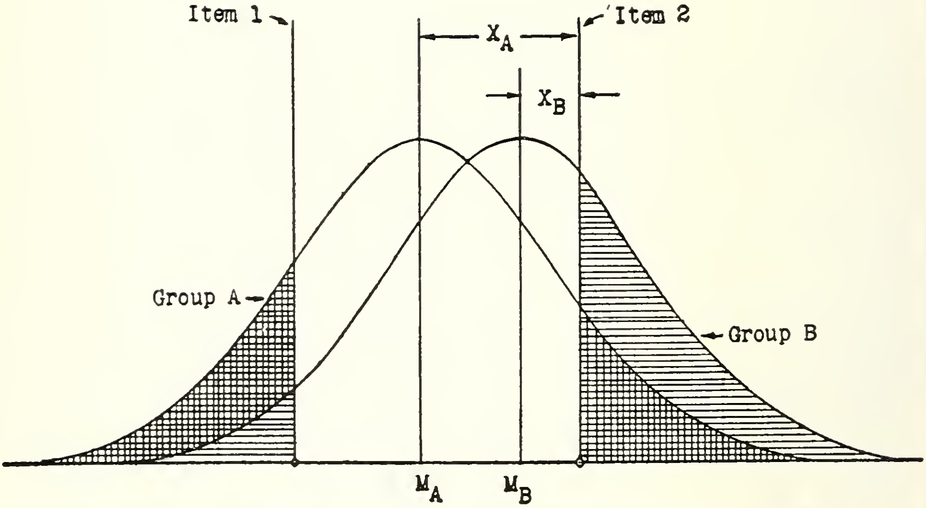


FIGURE 101

Correlation of Paired Sigma Values of Binet Items

(Thurstone, 1925. By permission of the Editor, *Journal of Education Psychology.*)



Item 1
X_A
Item 2
X_B
Group A
Group B
M_A
M_B

FIGURE 102. **Scale Values of Two Test Items from the Distributions of Two Groups of Individuals**

(After Thurstone, 1925. By permission of the Editor, *Journal of Educational Psychology.*)

the other. Figure 102 represents two such groups assuming the common conditions, i.e., a normal distribution and unequal sigmas. The per cents of each group passing Item 1 relate both groups at one point on the scale. Similarly, the per cents passing Item 2 orient both groups at another point. If fifty items are used, the means (M) of these SD scale values and the sigma (S) are determined with considerable accuracy.

4. Find the relation between these sigmas. This is done by dividing one by the other S_x/S_y. The results for the data in Table 113 are:

$$\frac{S_7}{S_8} = \frac{1.48}{1.32} = 1.12$$

This indicates the relative dispersion of the two groups when SD scaling has been applied.

In scaling for several groups, the sigmas of one of them, generally the lowest, or the most representative of a true sample, is taken arbitrarily as the unit, and the others are calculated in proportion. Thus, in Table 113, Column 2, the sigmas of age groups are all scaled from that of the 3-year-group. The 4-year-group had a sigma 1.19 as large as the 3-year, etc.

TABLE 113

Absolute Means and Sigmas of Age Groups

	1	2	3	4
Age	*M*	*SD*	M_T*	*CV*($\frac{2}{3}$)
$3\frac{1}{2}$	0.00	1.00	11.00	.91
$4\frac{1}{2}$	1.23	1.19	11.23	1.05
$5\frac{1}{2}$	2.34	1.26	13.34	.94
$6\frac{1}{2}$	3.37	1.26	14.37	.87
$7\frac{1}{2}$	4.06	1.33	15.06	.89
$8\frac{1}{2}$	4.87	1.49	15.87	.93
$9\frac{1}{2}$	5.43	1.50	16.43	.91
$10\frac{1}{2}$	6.19	1.45	17.17	.82
$11\frac{1}{2}$	6.77	1.49	17.77	.84
$12\frac{1}{2}$	7.31	1.50	18.31	.82
$13\frac{1}{2}$	8.04	1.53	19.04	.79
$14\frac{1}{2}$	8.68	1.79	19.68	.91

* M_T is the mean when zero is − 11 points below the mean of the $3\frac{1}{2}$-year-group. The use of this developmental zero makes the CV's in column 4 approximately equal. Columns 3 and 4 were added by the writer.

(After Thurstone, 1925, Table 3. By permission of the Editor, *Journal of Educational Psychology*.)

5. Find the relation between the means of the groups. This can be done from the regression of the SD scale values, if we

assume some arbitrary point for one of the means, since we have found the sigmas and the correlation coefficients between adjacent age group scale values of items. In Table 113, Column 1, the 3-year-group mean is arbitrarily taken to be zero and the 4-year-group was found to be 1.23 points higher when the 3-year-group sigma is used as a unit. Any mean or any other related point can be used for the arbitrary zero. These choices are purely for convenience since relationships are not changed by changing the zeros.

6. Find the SD values for each item. If the original sample is adequate, the scale value of any item is equal to the mean of any age using the item plus the product of the sigmas value of item and the sigma of the age group. Hence, for 7-year-olds:

$$\text{Scale value} = M_7 + X_7S_7$$

Thus, Item 42, Naming Months, was passed by .353 of the 7-year-olds, which gives it a sigma value in this age of +.38. Table 113 shows the mean and sigma of the 7-year-olds to be 4.06 and 1.33; hence,

$$\begin{aligned}\text{Scale value of Item 42} &= 4.06 + (.38 \times 1.33)\\ &= 4.57\end{aligned}$$

Approximately the same scale value could be secured from any age group that uses this item, but the values which are considered to be the most accurate are those secured from ages where the per cents passing are near 50 per cent. The reason is that random errors in measurements are more likely to occur at the extremes of a group where fewer cases are found. Thurstone suggests averaging two scale values by noting the two ages which give per cents passing the item just above and below 50 per cent. These per cents are changed to the corresponding SD values for each age group. Then the scale value of a test question which was common to 7- and 8-year-groups would equal:

$$\tfrac{1}{2}\left[(M_7 + X_7\sigma_7) + (M_8 + X_8\sigma_8)\right]$$

If this formula is applied to the data of Item 42, Naming Months, it appears that .353 of 7-year-olds and .616 of 8-year-olds passed the item, and that the SD values are +.38 and −.29, respectively. Table 113 shows the means and the sigmas of the 7- and 8-year-olds; hence,

$$\begin{aligned}\text{Scale value of Item 42} &= \tfrac{1}{2}\,[(4.06 + .38(1.33) + (4.87) + (-.29)(1.49)]\\ &= 4.50\end{aligned}$$

A. SD Scaling of Total Scores

In the same manner that individual items can be scaled, it is possible to scale total raw scores. The per cent of a group exceeding a particular score is found and changed into a scale value as noted above.

Although in rate tests, such as speed-of-tapping, where all items are much alike in difficulty and where the score is the number completed in a given time, SD scaling cannot be applied to single items, it is possible to scale total raw scores. In the case of information tests, when items are of different difficulties and time is not limited, SD scaling of total scores is not the best plan because two persons may receive the same total score by succeeding on different types of items. However, Richardson and Stokes (1935) scaled such tests by individual items and by total scores and secured very similar results.

The SD scaling of total scores was done extensively by McCall (1922), who urged that all test scores be transmuted to standard scores of a 12-year group. He called these T scores. A similar practice using Q as the unit has been applied by Van Wagenen (1934) to the *Unit Scales of Aptitude.* He called his scale a C Scale. Its units are each one tenth of a quartile deviation of a chronological age group. Thorndike (1926) scaled the CAVD Tests using the standard deviations of total scores of items which had been previously arranged in order of difficulty by a rating method. see Chapter 12. He also scaled total raw scores for a number of standard group tests including the United States Army Alpha, the Otis Advanced, and the Terman Group Test. Thurstone and Ackerson (1929) scaled MA's of the Stanford-Binet using a below-average group. Following Thurstone's method, Odom (1929) scaled the results of six well-known group tests, and Greene (1932) scaled eight nonverbal tests. Such scaling activity should lead one to consider both its limitations and the valuable results.

IV. LIMITATIONS OF THE METHOD

Scaling techniques which determine SD scales values from difficulty, as shown by the per cent of persons in a group who pass each item or score, have several serious limitations. In fairness to this method it should be stated that the following limitations are common to nearly all scaling methods:

1. The scaling does not show the nature of the difficulty. It may be due to a number of causes, such as rarity of the item, complexity, lack of motivation, or social inhibitions. In many instances it is important to know what kind of difficulty is present. This information can only be known by careful analyses of processes of individuals in the sample group.
2. Item difficulty is dependent upon a group of persons measured in a particular environment. The probability is small that items will retain exactly the same relative positions when another group of persons from a somewhat different environment is tested. Just how much variation can occur will be known only from continued research.
3. The influence of the position of an item in a test may be important, as the other items may form a context which influences difficulty.
4. The scores of the groups to be compared must be representable on the same base line. When a number of different factors are important in a test, this assumption may be inaccurate.
5. The implication is that all items will show similar growth curves. Table 113 shows this to be true with small variations for some Binet items. Figure 34, however, shows a great deal of variation in curves for the growth of knowledge of particular words.
6. The assumption is made that all groups compared have a normal distribution of scores; hence the method may only be applied where normalcy has been demonstrated.

V. SOME RESULTS OF SD SCALING

The main advantage of SD scaling is that it results in a scale which is independent of the raw-score units, and which is entirely dependent upon measured differences in the behavior of a known group of persons. SD units are equal in the sense that they are equally-often-noticed differences, or equal by statistical determination.

A number of studies have produced exceedingly interesting results concerning (1) selection of items, (2) age group dispersions, (3) true zeros, (4) mean growth curves, and (5) natural laws.

A. Selection of Items

A selection of items without regard to their scale values may lead to a bunching of items at various levels and a scarcity of items at various other levels. This situation is shown in Figure 103, which indicates the results of Thurstone's (1925) scaling of Binet items from the 1916 Stanford Revision. The bunching is marked between the points 2 and 3, 6 and 7, which correspond to 5- and 6-year-levels. On the whole, however, this figure shows a fairly even distribution of difficulty among the items.

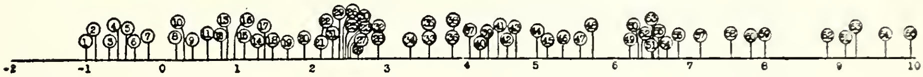


FIGURE 103. **Burt's Binet Test Items Scaled in SD Units**

(Thurstone, 1925. By permission of the Editor, *Journal of Educational Psychology*.)

B. Age Group Dispersions

The dispersions of age groups as shown by raw scores were found by a number of persons to be nearly the same over wide age ranges. In developing measuring scales, some have assumed this to be the true situation. The work of Thurstone (1929), however, shows that when an SD scale is used, the standard deviations of age groups increase with age. The increase is due to the fact that raw scores give equal weight to all items, whereas SD scaling gives more weight to harder items.

C. True Zeros

The calculation of age group means and sigmas on SD scales has led to the calculation of zeros of development. Thurstone (1928) found that on certain tests standard deviations had a linear relationship with means for different age groups. A convenient and usual way of stating this fact is to say that the coefficient of variation would be constant, if measured from absolute zero, see Table 113, Column 4. If we assume that the factors which determine the known part of the growth curve also control the unknown part, then, Thurstone argues, the true zero of ability will be found where there is just no variability in the group. This true zero may be calculated, or it may be indicated graphically, as in Figure 104, where sigmas are plotted against means. By extending the line of best fit to a point where it cuts the base line, the true zero may be found. When this technique is used, the true zero for Woody's data appears at approximately −3 sigmas, for the Illinois, −5 sigmas, for the Otis Advanced, −11 sigmas, and for Dearborn I, −8 sigmas. The reasons for these differences in true zeros are not easy to ascertain. Two hypotheses are put forward to explain why the CV's shown in Figure 104 vary. One is that tests involving

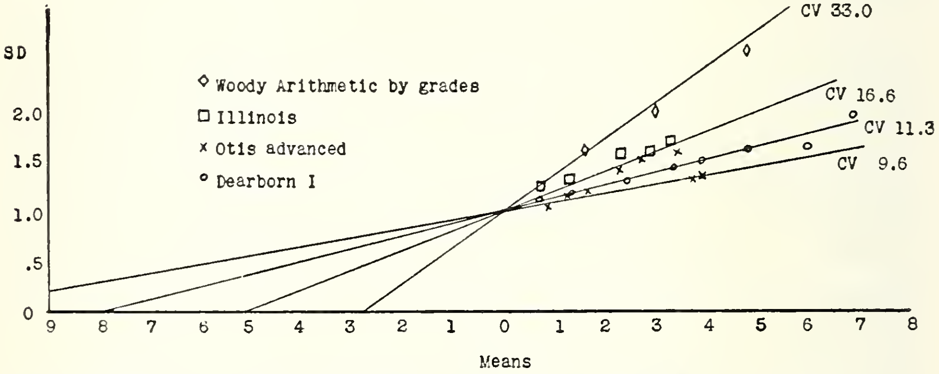


FIGURE 104. **Linear Relationship between Means and Sigmas**

(Drawn from data presented by Odom, 1929; Thurstone, 1925, 1928; and Richardson and Stokes, 1935.)

abstract inferences may have greater CV's than tests which have large specific factors, such as rote memory and motor coordination. A second is that variations in populations selected have caused smaller SD's at certain ages. Richardson and Stokes (1935) have pointed out the similarity of true zeros for group tests of general intelligence, and have suggested that there is a universal pattern involved here, with true zeros approximately 6 sigma below the means of any age group, and therefore a CV of 16.8. More research is needed to indicate the approximate zeros from normally selected groups for various tested abilities. The discrepancies of the present figures give promise that the method is valuable for precise analyses.

Terman and Merrill (1936) have reported that the SD's for IQ's remain fairly constant, being approximately 16.6 for the various age groups. They intentionally produced this result above the 13-year-old group, but IQ's below this group are calculated by dividing the mental age by the chronological age. The MA's of a particular age will have the same coefficients of variation as the IQ's because dividing numbers by a given number does not change their relationships. The MA Scale is a rough form of absolute scaling because the items are chosen so as to make the average MA equal the average CA of a group. This procedure results in selecting items which are passed by a particular per cent of the group. Since the mean growth curve from 6 to 12 years is probably nearly straight, the items are fairly well selected to give equal-appearing intervals. The CV's of MA's and of IQ's, therefore, correspond very closely to those found by SD scaling methods for similar types of items and groups of persons.

D. Mean Age Curves

The results of SD scaling throw some new light on the nature of growth curves because the data are free from raw-score distortions.

Thurstone's (1925) scaling of Burt's data resulted in a growth curve, Figure 105, which is negatively accelerated and has no inflection point from 3 to 13 years. Thurstone and Ackerson (1929) scaled MA's from the 1916 Stanford-Binet from a below-average sample of children in Chicago (see Figure 7). On the basis of this study an S-shaped curve appeared with an inflection point at approximately 11 years and some irregularities after 13 years. They believed that the inflection point would be earlier for normal or advanced groups. Inspection of the original data, however, indicates that the inflection point may be entirely due to the selection of atypical cases in the clinic.

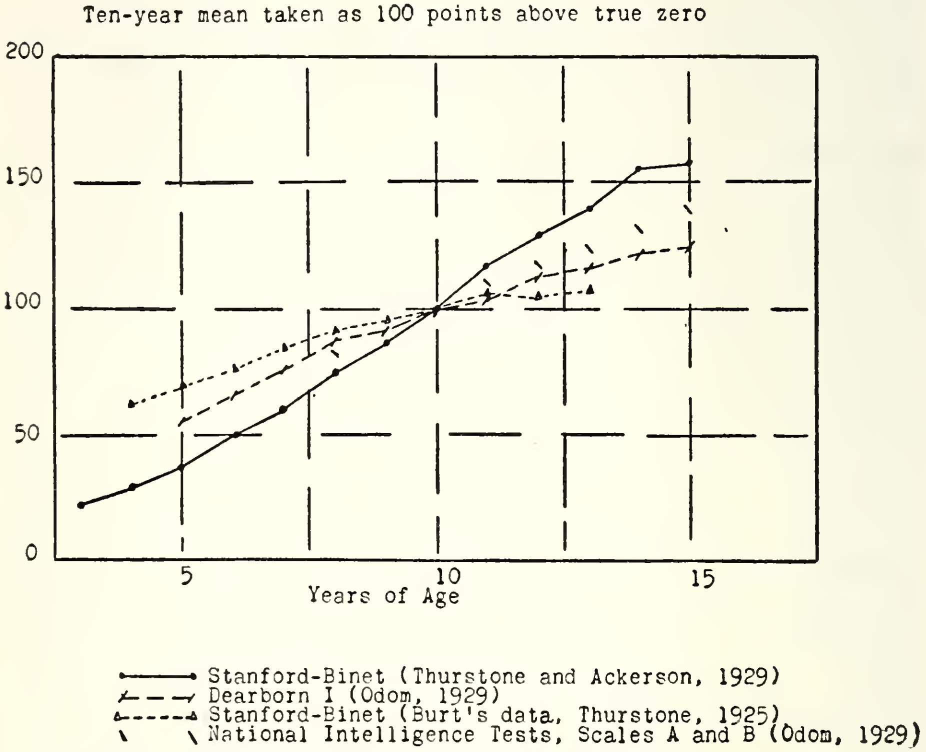
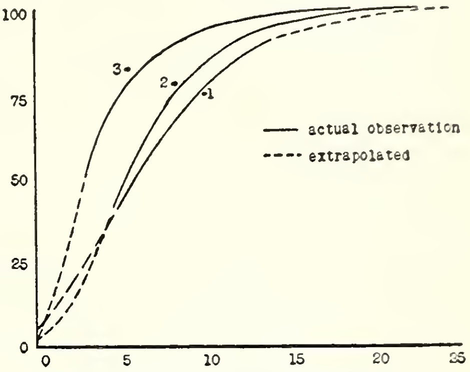


FIGURE 105. **Absolute Scaling of Mean Growth Curves**

Age-group-mean curves were calculated by Odom (1929) on six group intelligence tests using 87,486 cases. His results, some of which are shown in Figure 105, all give negatively accelerated curves with no inflection point, even though some groups as young as 5 years are included.

Richardson and Stokes (1935) found a very similar negatively accelerated curve, which is drawn in Figure 106 to show a possible relationship with maximum development. They found that their growth curve fitted the growth and practice curve suggested by Courtis (1929). Assuming that the same factors which govern the known part of the curve also control the unknown part, they extrapolated the curve to birth and to 25 years. On this basis, the average person is approximately 61½ per cent of his maximum at birth, 51 per cent at 6 years, 77 per cent at 10 years, and 92 per cent at 15 years. These figures would apply, with very small errors, to all the data scaled in Figure 105, with the exception of the curves from Binet scores of Thurstone and Ackerson (1929), which seem to be atypical.



(100 points above true zero is placed arbitrarily at 20 years.)

FIGURE 106

Mean Percentage Growth

(1. from Richardson and Stokes, 1935; 2. from Greene, 1931; 3. from Thorndike, 1926, p. 463.)

Figure 106 shows three curves, all of which come from large, fairly random samples of populations scaled in SD units so that the age of 20 years has a mean of 100 points above a calculated zero. Curve 1 is that of Richardson and Stokes' data from a mixed verbal intelligence test. Curve 2 fits very closely the means for aiming, tapping, feature discrimination, and pencil mazes. Both of these curves were scaled to absolute zero according to Thurstone's (1929) method. Curve 3 shows Thorndike's (1926) data for the CAVD tests. In this case the "true" zero was estimated by a rating technique below the age of 3 years, combined with an SD scaling technique above the age of 3 years. This curve is, therefore, not comparable in scaling with the other two. Since, however, the shape of the measured parts of all three curves is very similar, it is likely that with similar scaling they would be very similar in slope.

Curves 1 and 2 have inflection points near the fourth year of growth. The inflection point in the curve for nonverbal skills appears a little earlier than the curve for verbal and number skills. This may be an artifact of the scales, or it may represent a real difference in the developmental rates of these skills in these persons. More research is necessary to indicate the causes of these differences. Studies may show that some of the assumptions are unnecessary or partly in error, or that some other curves will represent the data more adequately. The work reported here presents a plausible set of assumptions, and these curves give bases of comparison which are relatively free from raw-score distortions.

E. Natural Laws

Another interesting application of scaling is seen in the field of psychophysics. Woodrow (1937) found that SD scaling of reaction times to lights of different intensities and with different preparatory intervals yielded a definite law of the relationship of effects of various intervals. He writes that . . . "entirely different results are obtained when raw scores are used in place of scaled scores, and that this law . . . could not have been discovered without the use of absolute scaling — a fact that suggests that psychology will not possess many broad quantitative laws until raw scores are more generally subjected to scaling techniques."

23

THE EVALUATION OF JUDGMENTS

I. THE PROBLEM

The preceding chapter described ways in which test items could be given absolute scale values and combined into scales which have equivalent steps — equivalent, that is, in the sense that they represent differences in difficulty which have equal standard deviation values in a particular population. This technique may also be applied to preference tests when they are in the form of a series of items which are to be judged. Psychologically there appears to be no sharp line of separation between tests of skill or information and tests of preferences. For instance, a person's judgment that statements are *true* or *false* may be scored in the same way as his judgments of *like* or *dislike* for a list of recreations. The scoring key for a factual test is made from an agreement of experts concerning the truth of each statement. The person's score shows how well his answers agree with those of the experts. The scoring key for a scale of preference is likewise constructed to show how well an individual agrees with the preferences of a group selected for some particular quality.

.A problem arises in the scaling of preferences or comparisons when they are not given in the customary standard test form. Three other comparison methods are commonly found which will be described and evaluated: the paired-comparison method, the rank-order method, and rating methods. The scaling of results from these comparisons is based on the assumption of Cattell (1902) that equally-often-noticed differences may be defined as equal unless they are always or never noticed. This statement is called the *law of comparative judgments.*

II. PAIRED-COMPARISONS METHOD

A scale which has equal-appearing-intervals may be constructed by comparing pairs of items directly, and then arranging them in such a way that the proportion of judges placing one above the next will be the same. In this method, each item must be judged to be better or worse than each of the rest, using some order of comparison which will not suggest the answer. If ten items are to be compared, one must make $\frac{n(n-1)}{2} = \frac{10 \times 9}{2} = 45$ judgments. These 45 comparisons must be made by a number of persons, or by the same person a number of times. The more judges, the better will be the results, if all the judges are equally competent, because chance errors tend to become less important as the number of judgments increases.

Fechner (1871) suggested a method for changing per cents to standard deviations of a normal curve. Witmer (1894) and Cohn (1894) used it to investigate aesthetic responses to form and to color. Titchener (1902) used it in studies of pitch and rhythm discrimination. A thorough mathematical treatment for the scaling of paired comparisons was given by Thurstone (1927), who took into consideration the various assumptions which must be made in calculating scale values. These assumptions and calculations cannot be described in this book, but the simplest case assumes that (*a*) the distribution of responses by a large number of judges to any one stimulus will be in the form of a normal curve, (*b*) the judgments of differences between two stimuli will also fall into a normal distribution, and (*c*) that errors made in responding to one stimulus are not correlated to errors made in judging the second stimulus.

When these assumptions can be made, and when judgments have been tabulated, the data may be arranged as in Table 114. Here the results of comparing six photographs for excellence in composition are shown by the per cents of judges who thought that one photograph was better than each of the others. This table shows that 70 per cent of the judges believed that A was superior to C, etc. The scale values for each item can be found by changing the per cents into standard scores, from Table 13, page 60, and then combining them. One should consult Thurstone (1928) or Guilford (1936) for examples of this technique.

TABLE 114

The Paired-Comparisons Method

Item	*Per cents preferring each item*					
	A	B	C	D	E	F
A		.30	.24	.15	.10	.01
B	.70		.37	.22	.10	.15
C	.76	.63		.33	.05	.08
D	.85	.78	.67		.42	.33
E	.90	.90	.95	.58		.39
F	.99	.85	.92	.67	.61	

NOTE: 70 per cent of judgments preferred A over B; 76 per cent A over C; etc.

An inspection of this table shows a marked tendency for the judges to be fairly consistent, but there are several discrepancies. For instance, Item E is judged superior to F, when both are compared to B. Such discrepancies are indications that the scale does not have high internal consistency. The judges probably select different aspects of a photograph for comparisons on different occasions.

Although the paired comparisons method is usually considered a very precise way of securing judgments, it is seldom used because simpler methods seem to be adequate.

III. RANK–ORDER METHOD

When the number of items to be scaled is more than 10 and the number of judges is large, the paired-comparison method becomes very laborious. The rank-order method overcomes this difficulty without much loss of effectiveness. It is far easier to rank 30 items than to make comparisons of 3,045 pairs. The rank-order method was used by Cattell (1903) in studying American scientists, and Wells (1908) in evaluating ten traits of leading American writers. Hollingworth (1911) and Strong (1911) applied it in studies of advertising appeal and memory value.

TABLE 115

Ranks Assigned to Samples

Item No.	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*Md.*	*Centiles* *	*SD*
				Judges						
1	2	1	3	1	2	1	1	1	16	41
2	3	2	5	3	1	3	3	3	50	50
3	4	4	4	4	4	4	4	4	66	54
4	1	3	2	2	3	2	2	2	33	46
5	5	6	1	5	5	6	5	5	83	59
6	6	5	6	6	6	5	6	6	100	78†

* There is an error introduced here because of the small number of items used. The centiles are all too high because they represent the upper limit of the ranks of the items. No simple way of correcting this error is available, but when 30 or more items are used the error is very small.

† This is an approximation. The items at the extreme of a scale are likely to be determined with considerable variation of SD values.

In computing scale values by this method, it is necessary to have the items ranked by a number of judges or by one judge a number of times. The median rank for each item is found rather than the mean, since the median is not influenced as much as the mean by extreme scores which sometimes are the result of errors. A median rank for any item is found by arranging all of its ranks in order of size, and then selecting the middle one. For instance, in Table 115 the first item has the following ranks 1, 1, 1, 1, 2, 2, 3, and hence a median rank of 1.

There are several procedures for changing median ranks into scale values. One of these proposed by Hull (1928) assumes that the items to be ranked have a normal frequency distribution. This assumption is probably true enough when the items are a large number of biological phenomena selected at random. If one had 100 items to rank, each rank would be a centile, since it shows the per cent of items which fall below a particular item. When any number of items are used, the ranks of the items can be changed from Table 13, page 60 into standard scores. The lowest item may be taken as the arbitrary zero point and the other items assigned scale values from this zero according to standard differences.

Two other important methods for scaling ranked items do not assume that the items are normally distributed, but that they may have any form of distribution. In one of these, Guilford (1936), calculated the per cent of judgments which place an item above a *composite standard.* These per cents are changed to standard deviations or standard scores as before. In the other, Thurstone (1931) calculated the proportion of judgments which place each item above or below every other item. This method converts the ranks into paired-comparisons and follows the procedure described for that method.

A variation of the rank-order method called the method of equal-appearing-intervals is useful when a large number of items are to be ranked. Instead of ranking them all in order, the items are placed in piles which appear to be equally spaced. Table 116 illustrates the procedure for 5 items which were arranged in 5 piles by twenty-four judges. More items were used but only 5 are shown since they are enough to illustrate the method. The median value of each item is preferred to the mean, for the items on the end

TABLE 116

Scaling by the Method of Equal-Appearing-Intervals

Item	*Piles*					*No. of Judges*	*Median Values*	*Q*
	1	*2*	*3*	*4*	*5*			
A	18	4	2	—	—	24	1.16	.33
D	4	10	8	2	—	24	2.30	.52
G	2	5	9	6	2	24	3.40	.72
N			6	10	8	24	4.18	.62
T				4	20	24	4.90	.27

NOTE: Item A was placed in the first pile by 18 judges, in the second pile by 4 judges, and in the third pile by 2 judges, etc.

piles will not be normally distributed. The *Q* of each item shows the range of the middle 50 per cent of judgments. In Table 116 the ranges are greater for items which appear in the middle piles most frequently. This situation is commonly found, and it is due in part to the fact that the scale does not extend far enough to give normal distribution for the highest and lowest items. The *Q* is thus reduced in some items by their extreme position. The *Q* is also an indication of the inability of judges to agree upon the relative position of an item. There may be disagreement owing to random errors in discrimination or to different standards of excellence. A large *Q* may thus be an indication of chance errors, or a lack of internal consistency in the scale. In either case, the items with the smaller *Q*'s are preferred for scale construction.

Sanford (1908) described the method of equal-appearing-intervals in scaling weights of envelopes. Thorndike (1910) had forty judges sort 1,000 samples of penmanship into 11 piles. Hollingworth (1911) had 39 jokes classified in 10 degrees of humor. Hillegas (1912) had judges classify English compositions into classes. Thurstone (1928) applied this method to establish attitude scales toward a number of political or social issues such as the established church, Table 7. Approximately 300 statements which voiced approval, indifference, or disapproval of the issue were sorted into 9 piles. The median value and the Q were found for each item. Finally, 15 statements were chosen to represent nearly equal steps of attitude ranging from very favorable to very unfavorable. Statements with the smallest Q's were preferred and two checks of validity, not described here, were applied. The care used by Thurstone and his students in the construction of approximately 30 different scales has made them among the most highly valued.

Remmers and his students (1934) showed that individuals made nearly the same scores when a group of items were presented either in a haphazard order or in the order of their scale values. The latter allowed a much more rapid scoring than the former. They also pointed out that a great economy in attitude scale construction could be effected if a general scale of opinion toward a class of social phenomena were devised. They forthwith published six generalized scales of attitudes toward:

1. An institution, such as war or Sunday observance
2. A race, such as the Chinese or the Negro
3. A homemaking activity, such as child care or preparation of a meal
4. A moral or social practice, such as petting or drinking
5. An occupation
6. A school subject

These generalized scales were composed by following the procedure devised by Thurstone. The correlation between the generalized Form A, Table 117, and Thurstone's Scale of Attitude Toward Communism was .816, and between the generalized Form A and A Scale of Sunday Observance, .83. Correlations between the generalized Form B and Thurstone's Scale of Attitudes Toward the Negro was .669 and Toward the Chinese, .72. These figures doubtless indicate similar attitudes in both scales, although the generalized forms usually contain a few items that are not well adapted to some of the situations in which they may be used. The generalized forms have also been criticized by Stagner and

TABLE 117

Attitude toward Any Institution

Scale Value	*Q Value*	*Item No.*	*Item*
11.1	0.8	2	It is the most admirable of institutions.
10.2	1.6	9	It is a strong influence for right living.
9.1	2.5	17	It is necessary to society as organized.
8.2	2.0	20	It does more good than harm.
7.4	3.2	21	It will not harm anybody.
6.1	2.6	23	It is necessary only until a better can be found.
4.9	1.9	26	It does not consider individual differences.
4.4	2.1	29	It represents outgrown beliefs.
3.3	1.7	32	It is too selfish to benefit society.
2.8	1.9	36	It is hopelessly out of date.
1.9	1.3	42	It will destroy civilization if it is not radically changed.
1.7	1.0	44	It benefits no one.

(Remmers *et al.*, 1934, Samples from Form A, p. 19. By permission of the Editor, *Journal of Social Psychology.*)

Drought (1935) because they found that scale values changed considerably when the same form was used for different items.

The effectiveness of various methods of comparison has rarely been studied, but Hevner (1930) scaled samples of handwriting by three methods: paired-comparison, ranking, and equal-appearing-intervals. She found that the scale values from the first two methods were very similar to each other but different from the third. The third gave less precise discrimination among the better samples of handwriting. On the other hand Ferguson (1939) reported very close agreement between scale values from the three methods, and he concluded that the equal-appearing-interval method was superior in both economy and accuracy. In this method the scale value of an item was not greatly affected by the inclusion or exclusion of other items. Aesthetic judgment, according to Bullough (1908), is disturbed by paired-comparisons. Conklin and Sutherland (1923) found that consistency in judging the humor of jokes was less when a ranking method was used than when a rating scale was employed.

IV. RATING-SCALE METHODS

Rating methods use scales which may have been established by one of the methods just described, or which may be quite arbitrary. The rater is asked to indicate on the scale the position of each person or item to be rated. This procedure is less time consuming than ranking items and more interesting to many judges. It is very widely used for evaluating beliefs, preferences, and personal traits.

A. Forms of Rating Scales

The form in which a rating scale is finally cast is of importance because it determines to some extent the accuracy and the speed with which a rating can be made. Four common forms will be described.

1. A Classified Form. This form asks the rater to indicate his judgment by marking the name of a class. A very common classification uses the headings *excellent, good, fair,* and *poor*. Initials or numbers are often used to designate these headings, as in Table 118.

TABLE 118

Rating Scale for Items concerning the Value of Music

As a first step in making this scale we want a number of persons to rate these statements by assigning them to nine different classes. We will call these classes *A*, *B*, *C*, *D*, *E*, *F*, *G*, *H*, and *I*. If you find a statement which you believe expresses the highest appreciation of the value of music, underline the letter *A*. For a statement which seems neutral or non-committal, underline *E* (the middle letter) while for those statements which express the strongest depreciation of music, underline *I*. Other degrees of appreciation or depreciation may be indicated by underlining one of the intermediate letters.

A B C D E F G H I 23. Music stimulates and encourages me in my life work.
A B C D E F G H I 24. To me music is of no greater or lesser importance than any other of the arts and sciences.

(From Seashore and Hevner, 1933, p. 369. By permission of the Editor, *Journal of Social Psychology*.)

Another classification uses statistical terms, such as *above average*, *average*, and *below average*. Sometimes a rater is asked to classify persons into quarters or fifths or thirds of a group.

In rating one's preferences the words *like, indifferent,* and *dislike* are often used, as in Table 79, page 461. Sometimes one simply classifies the item using one of two headings, such as *present* or *absent, yes* or *no, like* or *dislike, agree* or *disagree.*

Classifications are often combined with other rating methods, as shown in Table 91, page 502. For purposes of summary and comparison, classified ratings are usually given numerical scores, as in Table 89, page 495.

2. A Descriptive Form. In this form each step in the scale is designated by a word or phrase which describes particular behavior patterns as in Table 119. Sometimes these descriptions are very elaborate.

3. A Graphic Form. Here a straight line is provided, often with numbered spaces as in Table 119, and the rating is made by placing a mark on the line to designate one's judgment. This graphic form is rarely used by itself, but

TABLE 119

Distribution of Ratings of Nurses: University of Michigan School of Nursing

Rating of Performance in Service

Directions:			Item # 5			*M.D.*	*Q*
Adjustment to situations:	Sometimes at a loss in familiar situations	Slow to adapt to new situations	Learns new arrangements fairly soon	Quick to adjust to new routine	Very quick to respond to emergencies		
			Total				
1st Rater:	10	20	15	3	2	2.27	.67
2nd Rater:		9	32	9		3.01	.37

Note: When the medians and quartile ranges were calculated, each space was alloted one point on a scale from 1 to 5. The lowest value was assigned to the space at the left, and the highest to that at the right. When the values were interpolated, the whole points were assumed to be located in the middle of the space.

it is usually combined with a classified or descriptive form.

4. A MAN-TO-MAN FORM. This form, Table 120, was designed to make comparisons more definite by placing names of men known to the rater in blanks on a key sheet to represent standard levels of ability. Each person to be rated was compared to this standard and assigned a numerical score. In 1918 this man-to-man scale was used for rating a large number of United States Army officers.

B. Advantages of Various Forms

Since each form of rating has both advantages and disadvantages, it is hard to find clear evidence showing that one is better than another. Reported results of the direct comparison of two forms are very rare. Several experimenters have tried various forms, one after another, without finding any large differences in reliability or validity, see Chapter 16. However, when different forms are designed to have about the same function, they differ in a number of aspects, such as:

1. *Printed space:* Classified ratings such as those in Tables 79, 89, and 91 undoubtedly require less space than the other sorts.
2. *Ease of construction:* The descriptive and the man-to-man forms require more labor in their construction than the others.
3. *Ease for administration:* If one considers under this heading the time needed to explain the form and to mark the rating, the forms under consideration are nearly alike. except one, which is more time consuming. This one is a statistical classification form which requires that only a certain proportion of a group be given a particular rating. The necessary procedure requires first a ranking. then a grouping, and finally a rating according to the prescribed proportions.
4. *Time needed for scoring:* The forms which use numbers, can be most rapidly tabulated. They can also be

TABLE 120

United States Army Rating Scale

I. PHYSICAL QUALITIES Physique, bearing, neatness, voice, energy, endurance. Consider how he impresses his command in these respects.	Highest	15
	High	12
	Middle	9
	Low	6
	Lowest	3
II. INTELLIGENCE Accuracy, ease in learning; ability to grasp quickly the point of view of commanding officer, to issue clear and intelligent orders, to estimate a new situation, and to arrive at a sensible decision in a crisis.	Highest	15
	High	12
	Middle	9
	Low	6
	Lowest	3
III. LEADERSHIP Initiative, force, self-reliance, decisiveness, tact, ability to inspire men and to command their obedience, loyalty, and co-operation.	Highest	15
	High	12
	Middle	9
	Low	6
	Lowest	3
IV. PERSONAL QUALITIES Industry, dependability, loyalty; readiness to shoulder responsibility for his own acts; freedom from conceit and selfishness; readiness and ability to co-operate.	Highest	15
	High	12
	Middle	9
	Low	6
	Lowest	3
V. GENERAL VALUE TO THE SERVICE Professional knowledge, skill and experience; success as administrator and instructor; ability to get results.	Highest	15
	High	12
	Middle	9
	Low	6
	Lowest	3

(Note from p. 259 of *The Personnel Manual.*) It will be noted that it is really five separate scales, one each for each of the five essential qualities of an officer, namely, physical qualities, intelligence, leadership, personal qualities, and general value to the service. Each of the spaces is to be filled with the name of an officer who is taken as a standard for the qualification and the degree of the qualification indicated by the terms, "highest," "high," "middle," "low," and "lowest."

Each of the officers is ordinarily of the same rank as the rater and hence the rank next superior to that of the officer to be rated. Each of them is well known to the rater and stands in his mind as an exemplar of the qualification. With each of them he compares the officer to be rated on a man-to-man basis to find which one he most nearly equals in that qualification. The officer to be rated is compared with officers of superior rank because the object is to discover his fitness for promotion.

The accuracy of the result depends largely upon the care with which the rating scale is constructed. When instructions are followed closely and raters do their work conscientiously the ratings show a high degree of accuracy and uniformity.

(*The Personnel Manual*, Committee on Classification of Personnel, Adjutant General's Dept. 1919, p. 260. By permission of the Govt. Print. Office, Washington, D. C.)

adapted to machine tabulation. Forms which used descriptive phrases or graphic schemes require more time because some numerical score must be imposed on the results.

C. Relative Intangibility of Items

Regardless of the form used for rating, some items appear to be more difficult to discriminate than others. In physical judgments the finer discriminations are obviously the more difficult. Among personal traits, those that are complex, that have been ambiguously defined and that are rarely observed, show the most variations in results. This fact is illustrated in Figure 107, which shows the ratings on one boy on ten traits by from 10 to 15 observers. The observers were very consistent with one another in rating Resistance to Authority, Self-Assertion, Social Responsibility and Popularity. Less uniformity is indicated by the longer lines in the figure for the ratings of Poise and Leadership. Great lack of agreement is shown in ratings of Interest in the Opposite Sex, Awareness of Audience, Appearance, and Grooming Activity. The last two items, which one would think

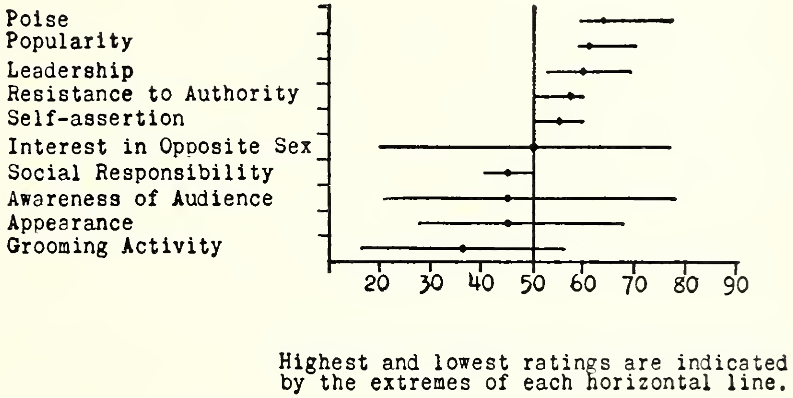


FIGURE 107. **Consistency of Ratings of Different Traits by Various Observers**

(Drawn from data in the files of the Institute of Child Welfare, University of California. Courtesy of H. E. Jones.)

could be judged uniformly, showed more variations than such intangible items as leadership and poise. These results should make one realize the need of a check on the inconsistencies of raters in any situation.

D. Number of Desirable Steps

The number of steps in rating scales range from two to a hundred or more. Four or five steps were most frequently used in ratings of attitudes and traits in 52 samples which were collected by the writer. In the self-rating of academic or vocational preferences three or four steps or classes are most common, as in Table 79. The best number of steps to use is that number which can be clearly distinguished by the judges in a reasonable time, and which does not distort the results. This may be determined by trying out scales with various numbers of steps. Symonds (1924) (1931) found from empirical evidence, that a seven-point descriptive scale was a little more reliable than a five-point scale when the judges were interested and the trait was clearly defined, such as neatness. When the trait was more vague, such as tact, or when the judges were immature or not very much interested, only five or four points were clearly distinguished.

Champney and Marshall (1939) compared two procedures in scoring a seven-centimeter graphic scale of home conditions. In one procedure a seven-point scale was used, one point for each centimeter. In the other procedure a seventy-point scale was used, one point for each millimeter. Both procedures were applied to two alternative forms applied within a three-week interval. The two forms yielded retest correlations of .665 for the seven-point scale and .766 for the seventy-point scale. This increase in retest reliability is more than could be expected by the usual methods of correcting for coarse scaling. It indicates that the judgments were more consistently recorded on the finer scale.

Another way of evaluating the number of steps to be used is to see whether or not all the judges use all the steps or bunch the cases in one or two steps. Judges have been found to differ a good deal in this respect. Thus Table 119 shows the distribution of ratings assigned to the same 50 nurses by two supervisors. The results show that the second supervisor used only three of the steps while the first used all five. If some steps in a scale are never, or almost never used, they probably have little effect on the scores. Occasionally, however, a lowest step which is never used causes the step next to it to be used more frequently than might otherwise be the case.

E. Halos and Logical Errors

Ratings on a number of traits by one person are subject to two sources of error which are sometimes rather important. One source is called a *halo effect,* and the other is a *logical error.* A halo effect is illustrated by a tendency to classify a person on the whole as good or bad, and then to rate him on all traits in keeping with this opinion. A logical error is introduced by a special interpretation of the task.

Halo effects and logical errors, of course, defeat one of the main purposes of rating scales, namely, to find out a person's relative strength and weakness. They can sometimes be avoided by rating all persons in a group on one trait at a time, instead of rating one person on all the traits listed before considering the next person. Another device for avoiding halo effects in ratings is an irregular arrangement of the steps or classes. A common arrangement of items is one in which all the most desirable characteristics are placed at the right-hand side of the page, the least desirable on the left, and the average traits in the middle. This arrangement may lead to a routine set of checks down the page by one who is too busy to use the rating carefully. If the most desirable traits are placed sometimes on the right and sometimes on the left, the rater is required to read the items more carefully. Mathews (1927) found that there was a considerable tendency in multiple-choice tests to check the responses printed near the left, rather than those printed near the right, when choices were printed across a page. When choices were printed in column, there was a marked tendency to check those nearer the top more frequently than those nearer the bottom of a series.

Halos are systematic errors which can be exposed by correlation techniques. An accurate rater will yield scores which correlate well with a true criterion and he will also be able to repeat his performance with a high self-consistency. A rater who depends upon chance will usually show a low correlation with the criterion and also a low self-consistency. A rater who is markedly biased will have a low correlation with the criterion but a high self-correlation. Adams (1930) has shown that persons often differ a great deal in their ability in estimating size of circles, lengths of lines, and also personal qualities of their friends. He found that a good judge of *self* tends to differ from a good judge of others, by being rated by his colleagues as happier, more sympathetic, generous, and courageous.

In relationships among personal traits, halo effects introduce errors which raise or lower correlations spuriously. When correlations between two dissimilar traits are as high as .90, halo effects may be suspected. The true relationships may be approximated by pooling the ratings of several persons on the assumption that personal prejudices will tend to cancel one another. This assumption brings up the interesting question of how many persons' ratings should be pooled to eliminate halo effects. Bradshaw (1930) found that the rating of a personality trait of students in college attained a correlation of .80 between two trials when at least five ratings were averaged. A reliability coefficient of .90 would be reached theoretically when the ratings by ten such persons were averaged. He reported many variations from these usual findings which depended upon the trait, the accuracy of the descriptions, and the raters.

A logical error is one in which traits are rated alike because the rater has some reason for believing that they are similar or occur together. He substitutes his logic for the direct observation of behavior. Newcomb (1931) found that ratings on a large number of personality traits in boys correlated much higher on the average, .493, than records of observed behavior of the same boys in the same traits, .141. He believed that the raters were not particularly prejudiced, but that they had accepted various theories and plans of personality which influenced their judgments.

F. Rules for Raters

From the considerations just discussed, the following rules should prove helpful in avoiding errors in rating procedures:

1. Each trait should be defined as clearly as possible, avoiding general terms.
2. Each trait should refer to only one relatively independent pattern of behavior — a well-nigh impossible accomplishment at present.
3. Raters should judge on the basis of their actual experiences, avoiding prearranged schemes which may not apply to a particular situation.
4. Complete integrity should be secured in rating either self or others. It may be an advantage to secure ratings from judges who are ignorant of their use.
5. Raters should check their definitions and scales of value with the accepted criteria frequently.
6. The combined ratings of several equally competent judges should be used rather than the ratings of one judge.

G. Combinations of Ratings

It has been shown that the most self-consistent ratings often come from combining the work of several raters. Raters often seem to cancel out one another's idiosyncrasies. The process of combining ratings is, however, often complicated by the fact that different raters vary in their leniency and in the scope of their judgments. This is well demonstrated by Table 119. The median ranks and dispersions are different for each rater. The first nurse ranked the group nearly a whole step lower than did the second nurse. This discrepancy means that the ranks assigned by one are not comparable to those assigned by the other. Such differences as these are frequently found among raters. Training tends to reduce the differences, particularly if judges are told to use all the classes and to make their results fall in a normal curve. There are, however, some objections to this procedure. It may lead to artificiality, and in many small samples the curve of distribution is probably not a normal curve.

In situations where it is desirable to have the results of all raters comparable, it is possible to change them all into standard scores. Symonds (1931, p. 81) has facilitated such changes by giving a table of the per cents which may be expected from 3- to 7-point scales, applied to two groups, one of 40 and one of 185 persons. However, Conrad (1932) (1932 A) found that adjustments to a common distribution did not significantly increase the validity of combined ratings in the United States Army Man-to-Man scale, Table 120, nor in the rating of intelligence of nursery-school children. He concluded that in these situations such time-consuming adjustments were not worth the trouble. Where very large differences occur between two raters, however, an adjustment to make their ratings comparable may be advisable.

H. Validity of Ratings

The desire to have some definite information concerning whether ratings reveal the true situation has led to considerable research along two main procedures. One of these correlates actual measures with ratings, and it is therefore limited to situations where actual measures are available. The other procedure sets up some combination of judgments as the nearest available approximation of the truth and compares other judgments with this.

An illustration of the first procedure is seen in the work of Marsh and Perrin (1925), who correlated ratings with various direct measures. The judges observed students performing certain tests, and then judged the performance. The correlation between estimated and measured aiming was only .36; between estimated and measured intelligence, .78; and between appraisals of card sorting, .68. Ratings of head size correlated .76 with measures of head circumference. These facts indicate considerable discrepancy and lead one to suspect that ratings of more intangible traits may be even less in keeping with the facts.

The second procedure is well illustrated by the work of Adams (1936), who also furnished formulas for showing certain qualities of a scale. In one of his experiments, on two different occasions a number of persons ranked ten printed circles in order of size. The two rankings of each person were correlated to give a *personal consistency*. These correlations were averaged to give *self-consistency* for the whole group. Another figure called *group-consistency* was secured by correlating the rank order assigned by each person with the rank orders of all other persons. In practice this procedure involved too many calculations; so a sample thought to be representative of all person-to-person correlations was taken. The self-consistency was found to be the same as the group-consistency in this experiment and also in a large number of other similar experiments with objects. Both group-consistency and self-consistency were high or low depending upon the difficulty of the discrimination.

In another experiment the same persons ranked ten students on personality traits such as courage. The results showed that self-consistency was considerably higher than group-consistency. This finding indicated that persons were more consistent with their own biases than with those of others. The group-consistency indicated the amount of random error, whereas the self-consistency was raised by systematic errors of personal sorts. Adams, therefore, proposed an index of *objectivity* found by dividing group consistency (GC) by self-consistency (SC). If GC equaled .81 and SC equaled .90, then the equation from which the index is derived will be as follows:

$$\frac{\text{GC}}{\text{SC}} = \frac{.81}{.90} = .90 = \text{objectivity index}$$

When this index is 1.00, group- and self-consistency are identical, as in judging the size of circles. The judgments are considered to be completely objective even though they may not be very accurate. Objectivity is therefore defined as the lack of systematic errors in judging. An objectivity index of less than 1.00 indicates the presence of systematic errors in the judge. The lower the index, the lower is the objectivity. This index is a very useful one which should be applied to both rating and measuring techniques.

In the experiment on judging sizes of circles there was an acceptable criterion for size, namely, the areas calculated from actual measures. When there are no widely acceptable criteria available, as in the judgment of courage, tact, or artistic ability, then there is no possibility of a conclusive check on the degree to which a rating approximates the truth. There are two procedures, however, which are designed to furnish criteria of approximate truth. In one, the central tendency of a group of judges is taken as the best approximation. In the other, the agreement of several similar procedures is taken as evidence that all are yielding true appraisals.

The use of a central tendency as the best approximation of a fact is of course open to question. If a number of judges make the same error, the opinion of a single expert will be closer to the truth. If, however, all judges have nearly equal ability, then a central tendency has been shown in many experiments to be nearer the fact than the judgment of any one of the judges. This discussion raises the very interesting question of who is the best judge, or, in the case of instruments or tests, of which is the least subject to errors. In some instances the qualifications of the judges or of the persons who constructed the tests are considered. Thus an expert in chemistry would be considered a more competent judge of chemical analysis than an untrained person. In many other cases, however, no accepted qualifications are available. Although some work has been done, further intensive research is needed to show the extent to which accuracy of a particular judgment is related to personal characteristics. Hollingworth (1922), Hoffman (1923), Shen (1925), Adams (1927), Conrad (1932), and others have reported that self-ratings are usually too high on desirable traits and too low on undesirable traits, and that superior individuals often underestimate themselves and the inferior overrate themselves. Adams found that a good judge of self is likely to be more intelligent, observing, sympathetic, generous, courageous, and happier than a good judge of others. Hollingworth believed that there was a positive correlation between possession of a desirable trait and ability to estimate it in others.

The other approach to the criterion for validity is made by using four ratings, all of which are designed to appraise the same quality. First, one must show that the appraisals are free from systematic error, by methods described for securing objectivity. Next, as Adams (1936) has shown by Spearman's logic the four appraisals may all be considered to measure the same function if together they satisfy the formula:

$$\frac{r_{ab}}{\sqrt{r_{aa}r_{bb}}} = 1.00$$

Here, if a correlation between raters *a* and *b* is equal to the square root of the product of their reliability coefficients, the validity index is 1.00. For example, if on two occasions persons A and B ranked ten photographs for their advertising appeal, and if the following correlations were found: r_{aa}, .60;

r_{bb}, .30; and r_{ab}, .43; then, by substitution of these values in the formula, we have:

$$\frac{.43}{\sqrt{.60 \times .30}} = 1.01$$

In this case Adams would conclude that both persons were using the same set of standards in rating the photographs, even though one rater was much more consistent than the other. If the index is much less than 1.00, the raters are not using the same standards.

24

MEASUREMENT OF NATIVE DIFFERENCES

The determination of the relative influence of environmental and inherited factors in the development of individuals is a time-honored problem which has attracted students of all the biological sciences. There are two general approaches to its solution. One approach is to note the effect of selective breeding on a number of generations; the other is to observe the results of various environments upon individual development. Both sorts of studies will be described, but for the most part the discussion will be limited to the consideration of human beings.[1]

I. SELECTIVE BREEDING

One method typical of the genetic approach to mental inheritance is to select a person who is judged to be either very superior or very inferior, and then to investigate his relatives in order to see how far they deviate from the norms for the whole population.

---

[1] The literature in the field of genetics is voluminous. Short statements of the problems, methods, and results of biological genetics may be found in Morgan (1928), Jennings (1930), Shull (1931), Newman *et al.* (1937), East (1936), Carmichael (1935), and Schwesinger (1933).

A. Relatives of Superior Persons

Investigations of superior persons are well illustrated by the work of Galton (1869), Cattell (1903), and Terman *et al.* (1925, 1930). Galton in his pioneer work investigated the relatives of 977 eminent persons. Each had achieved fame equal to that secured by only one person in about 4,000. He found 203 relatives who had achieved similar eminence, a fact which meant that the probability of having an eminent male relative was 134 times as great for an eminent man as for an average man.

Cattell (1903) reported an extensive study of the relatives and environments of 1,000 persons listed in *American Men of Science.* He found that both the births and the residences of these persons were concentrated in or near large cities where there were universities or rich scientific backgrounds. Massachusetts had produced relatively four times as many scientists as Pennsylvania, and fifty times as many as any southern state. In proportion to their numbers, parents who were in professions produced fourteen times as many scientists as commercial or trade groups, and twenty-eight times as many as agricultural groups. From these and other considerations Cattell concluded that the nature and degree of success were based on innate factors but that environment gave direction to the performance and often imposed enormous obstacles.

Terman's report was concerned with 643 children who had demonstrated by intelligence tests and school work a degree of mental ability reached by less than 1 in 200 children. He found among the eminent relatives of these children, six signers of the Declaration of Independence, two presidents and two vice-presidents of the United States, two Supreme Court justices, six well-known authors, two inventors, and many outstanding leaders in their communities. One family had 34 relatives who were named in Appleton's *Cyclopedia of American Biography,* or in *Who's Who in America.*

B. Relatives of Inferior Persons

Investigations of inferior persons are illustrated by Woodall's (1932) study of children of defective mothers. Of 385 children born to 192 mothers, 128 died at an early age. Of the remainder, 16 per cent were average or above average on mental tests, 40 per cent ranged from dull-normal to borderline, and 44 per cent were classed as mentally defective. These per cents should be compared with those in Table 58, page 290 to show marked differences from a normal IQ distribution. Other studies by Estabrooks (1915), Goddard (1914), and Davenport (1912) report the offspring during several generations of families that were impoverished and inferior.

C. Correlational Studies

Another genetic approach is the correlation of test scores or ratings of ability among related persons. Pearson (1904) obtained teachers' ratings for 2,000 pairs of siblings on eight traits; vivacity, self-assertiveness, introspection, popularity, conscientiousness, temper, ability, and handwriting. The ratings placed children in either two, three, or six classes. In spite of this rather coarse rating technique, the correlations were all very close to .51. The resemblances of pairs of brothers were nearly the same as those of pairs of sisters, and of brother and sister pairs. The ratings of ability yielded the lowest correlations, approximately .46; and conscientiousness and introspection the highest, approximately .60.

A little later Woods (1906) read biographical material concerning 671 members of royal families of Europe and rated each on a ten-point scale for intellectual ability and morals. His correlation between fathers and children was .30, between grandfathers and offspring, .16, for both intelligence and morals. Wood's report included many other points of comparison, such as individual interests and temperaments, which seemed to him to indicate that native factors were very important in the development of these traits.

After the introduction of standard tests of intelligence, a number of extensive reports of tested abilities of related persons appeared. An extensive summary of this material is given in the *27th* and the *39th Year Books of the National Society for the Study of Education.* The measurement and statistical techniques involved in these appraisals combine all of the hazards discussed in the Validation of Tests, Chapter 5, the Measurement of Practice, Chapter 20, and the Measurement of Growth, Chapter 21. Table 121 gives a composite picture of the results of several investigations. The striking similarity between the physical and mental data is apparent, and it leads to belief that physical and mental characteristics are inherited in the same manner. There also exists a hypothesis that the correlations for intellectual traits are too low. Thorndike (1928) found that the correlations between CAVD tests in siblings in high school were near .60 when corrected for attenuation.

TABLE 121

Correlations between Related Pairs of Persons, Northern European Stock

	Physical Measures	*Mental Measures*
A. Siblings, same home	.42–.53(3) *	.45–.55(1, 2, 4, 8)
B. Parent-child	.42–.53(3)	.45–.55(1, 2, 6, 8)
C. Parents	.00–.15(3)	.46–.49(1, 2, 6)
D. Midparent-Midchild	.64(3)	.69(1, 11)
E. Nonidentical Twins	.58(7, 7)	.53–.70(4, 5, 7, 8, 12)
F. Identical Twins	.91(7, 12)	.88–.97(4, 5, 7, 8, 12)
G. Cousins		.27(11)
H. Unrelated children at large		– .19–.09(9, 10, 8)

* These correlations are not strictly comparable since they were taken from different populations using different tests. The numbers in parentheses refer to the following articles:

1. Jones (1928)
2. Freeman *et al.* (1928)
3. Pearson (1915)
4. Merriman (1923)
5. Hirsh (1930)
6. Burks (1928)
7. Holzinger (1929)
8. Wingfield (1928)
9. Hildreth (1925)
10. Pintner (1931)
11. Dexter (1932)
12. Newman *et al.* (1937)

Since measures of physical traits are fairly precise, they would not be significantly raised by such corrections. The higher correlations of mental as compared to physical measures in related persons were believed to be due to environmental influences which affected the mental more than the physical traits. Hence the difference between approximately .50 and .60 was regarded as an indication of the effect of similar environment in reducing individual mental differences. The difference between correlations for siblings and for nonidentical twins was likewise approximately 10 points. In as much as these two groups are alike in inherited factors, the higher correlation of the nonidentical twins is thought to be due to a greater similarity in their environments than is typical of ordinary brothers and sisters.

D. Experimental Studies

The experimental control of native factors is well illustrated by the breeding of domestic plants and animals. A good example of breeding for mental ability was given by Tryon (1929), who tested white rats in a 19-choice maze. The animals which learned most quickly were mated with each other, and the poorest learners were likewise bred. The offspring of both groups were tested in the same maze and bred according to their abilities. This procedure was repeated to the eighth set of offspring. The result was a gradual separation into two groups of animals. The more able group finally consisted of rats all of which were superior to all but about 4 per cent of the less able group. The superior group, however, did not learn the maze any more quickly than the best learners of the original stock. This experiment seems to indicate that ability in maze learning is inherited, but that selection of superior animals did not lead to the development of abilities greater than those found in the original sample.

Although the degree of control illustrated in this experiment with rats has not as yet been reported for studies of human heredity, there is reason to assume that selective factors produce superior abilities. A persistent question is whether the level of intelligence of a community is rising or falling. Those who claim that mean ability is rising in a community usually point to the larger proportion of persons at present entering high schools and colleges, and to the fact that standards for entrance and graduation are apparently higher than those that formerly were in force. Those who think that mean intelligence is falling usually argue in much the same manner as a recent book by R. B. Cattell (1936a), which sets forth the results of surveys of 10-year-old children in typical urban and rural communities in Leicester and South Devonshire, England. The most emphasized finding is that mean IQ's of children in small families are higher than IQ's of children in large families. Cattell concluded very emphatically that the mean intelligence of these communities was gradually declining, and that dire economic and social results must follow. This conclusion, although conceivable, lacks proof along two lines: one is ignorance of differential death rates, and the other is the assumption that IQ's are almost entirely determined by heredity. This assumption does not seem justified either from the construction of the tests or from other studies reported below which indicate that environmental changes may bring about large alterations in test scores over a period of years.

In evaluating all of the studies of human beings mentioned thus far, it must be emphasized that environmental factors have in no instance been varied experimentally by the investigator. In a number of studies the environments have been reported in mere outline, and in all some evidence appears that superior environments were usually enjoyed by the persons who are rated or who prove themselves by tests to be the most able. Although it is widely believed that the more able persons aid in creating the more favorable environments, it also seems true that cultural traditions have often held back or stimulated maximum artistic and scientific self-development. One must therefore look further for research which holds native ability constant while the environments are varied, in order to note the effects of environment on test scores.

II. ENVIRONMENTAL CHANGES

The control of environmental factors is illustrated by studies of changes in home environment and in formal schooling. These two environmental forces undoubtedly react upon each other, but several studies will be reviewed which have emphasized one or the other.

In order to show the results of changed environments, it is desirable to have a control group representing growth in an environment which can be designated as normal. A report of children in fairly normal home situations by Jones and Bayley (1937) described 61 children who were repeatedly tested from birth to 6 years. Socio-economic factors showed nearly zero correlations with intelligence test scores before 2 years of age. Intelligence was found to be correlated approximately .50 with parent's education at all ages from 2 through 6. Intelligence scores showed low but gradually increasing correlations with father's occupation, family income, and a social rating of the home during this period.

A. Changes in Home Environment

Two studies of IQ's of foster children were published in 1928, one by Freeman and others in Illinois, and the other by Burks and others in California. The Illinois study followed the placement of 401 children in foster homes over a number of years. The California study was principally concerned with 261 children placed in foster homes. Both studies used the 1916 Stanford-Binet test and ratings of home environments. Ratings of home environments were found to correlate approximately .45 with IQ's of own children in both studies. The Illinois study found a similar correlation, .49, between home ratings and foster children's IQ's, but these figures for the California study were lower, .21 to .35.

Both studies reported that small increases in IQ's, from 5 to 7 points occurred when children were placed in homes that were rated considerably above average, after having been in inferior homes. The increases were larger for those who had been placed at earlier ages than for those placed at later ages, and for those who had longer residence in superior homes than for those with shorter residence. The Illinois study indicated that the placement in a poor or average home did not result in a significant IQ change. For a small group of siblings raised in different foster homes, the correlation of IQ's was only .25, and those in the better homes had somewhat higher IQ's. In spite of the fact that a large proportion of the parents were rated as mentally and morally defective, only 3.7 per cent of the children had IQ's below 70, and few cases of delinquency were found.

Neither of these studies can be called conclusive, for only small numbers of children were studied in the various subgroups, and in many cases there were no good indices of mental ability before placement in foster homes. It is possible that the more intelligent foster parents might have received an above-average group of foster children owing to various factors in selection.

The California study included an interesting statistical analysis of factors. The following correlations, corrected for attenuation, were found:

MA of parents and IQ of own children	.6036
MA of parents and Whittier Index of Environment	.7653
IQ of own children and Whittier Index	.4771
IQ of foster children and Whittier Index	.42

From these figures the percentage of the variance of own children's IQ's due to inheritance (parental IQ's) was calculated to be .33; the percentage due to environment .17; 40 per cent seemed to be due to hereditary traits from other ancestors, and approximately 5 to 10 per cent to random somatic effects of environment.

A more recent study by Skeels (1938) reported mental test scores of 147 children placed in foster homes under 6 months of age and tested 12 and 60 months after placement. The mean IQ's were found to be much higher than those of children of similar families who had not been placed in superior environments and the correlations between true mother's IQ and child's IQ was zero. The mother's mean IQ was approximately 84 and the children's 115.

Crissey (1937a) found that very young children lost about 10 points in IQ during residence in an orphanage for 18 months. He also reported (1937b) that the orphanage residence seemed to be more stimulating to older borderline children than a home for the feeble-minded, since children who were moved from the orphanage to the home lost in IQ, and those who were moved from the home to the orphanage gained in IQ. In another comparison, Crissey matched 50 adolescents in two orphanages with 50 in state schools for feeble-minded. Members of each matched pair were within 3 points of IQ and 6 months MA of each other at the time of the first test. Those in the orphanage gained 2.9 points in IQ and the other group lost 4.2 points during 6 months.

The effect of superior environments upon growth on intelligence test scores was reported by Skodak (1938), who studied 154 children placed in superior foster homes between the ages of birth and 6 months, and 65 children placed between 2½ and 5½ years of age. She found that the mean IQ's for both groups increased with the length of residence in the superior homes. There was a marked relation between mental level of foster children and occupations of the foster parents, but none between the children's IQ's and the mental level of their own parents. She concluded that the home was very important in setting the limits of mental development.

In a few reports little or no change in IQ is discovered after placement in foster homes which were superior in some ways to their original homes. Thus, Schott (1937) found that gains in IQ's made by children after placement were not unusual when compared to gains of children for a normal population.

Studies of Twins. A number of studies have attempted to appraise environmental effects by comparing twins raised apart. Merriman (1924), Hirsh (1930), Muller (1925), Newmann (1933), Lauterbach (1928), Wingfield (1928), McGraw (1935), and Mirenova (1935) have each reported a few cases. These are principally of interest for their detailed case studies. In many instances the tested differences seem to be related to differences in past experiences. Statistically, the number of cases was usualy too few to overcome typical errors of measurement. Moreover, the twin pairs have often not been separated for very long, or their separate environments have been fairly similar, or they have been separated relatively late in development.

One of the most extensive and thorough studies of the origin and resemblances of twins is that of Newman, Freeman, and Holzinger (1937). Fifty pairs of identical and fifty pairs of fraternal twins reared together were reported to show tentative norms. Also nineteen pairs of identical twins separated in infancy and reared apart were studied. Each person was measured by a battery of tests which occupied nearly two days. Among the physical measures were those of height, weight, head dimensions, hair color and pigmentation, handedness, palm and finger prints, and physical peculiarities. The mental tests included the Stanford-Binet, the Otis Self-Administering, the Stanford Achievement, the Downey Will-Temperament, the Woodworth-Mathews Personal Data, the Kent-Rosanoff Free Association, and samples of handwriting. Their findings, which are in harmony with nearly all the previous work, included the following:

1. Older pairs of identical twins, 13 years or more, showed slightly more resemblance than younger pairs, but older pairs

of fraternal twins were markedly less alike than younger pairs in mental and educational age. In height, no significant differences appeared, although the older groups had a little larger resemblance correlations.

2. Correlations of resemblances corrected for age and attenuation in physical measures were slightly larger than those for mental traits when twin environments were the same. Resemblances in motor tests and emotional traits were somewhat smaller than the rest. The quality of handwriting of identical twins was found to be more similar than that of fraternal twins, but not the general style or appearance of handwriting.

3. When twins were separated at an early age, resemblances in height and head dimensions were not affected by the greatest changes presented in the environment. Noticeable differences occurred however in weight, intelligence, and school achievement, which seem to be due to environmental differences. On the Binet and the Otis IQ's the identical twins who were reared apart showed nearly the same degree of resemblance as the fraternal twins reared together, less resemblance on the Stanford Achievement tests, and more resemblance on the Woodworth-Mathews Personal Data Sheet. Ratings of the relative degree of difference between environments of separated twins correlated with test results as follows: schooling with educational achievement, .91; schooling with intelligence tests, .46 to .79; social environment with intelligence tests, .32 to .53; physical environment and weight, .60.

4. Both groups of twins were found to have nearly normal ranges of intelligence and physical dimensions.

5. Approximately 80 per cent of the variance of physical measures among fraternal twins reared together was calculated to be determined by native factors, 70 per cent of the variance in mental measures, and 30 per cent of motor measures. In each case the remainder of the variance was allotted to environmental factors. For identical twins reared apart, however, the share of native factors in determining the variance in weight was calculated to be 13 per cent, in Binet IQ's, 49 per cent, and in the Stanford Achievement test, 13 per cent. These figures mean that when environments are very different, the share of environment in determining traits which are susceptible to environmental influence is large. When environmental differences are small, the variance in these traits must be largely due to heredity.

B. Special Training

A number of studies have attempted to appraise the effects of special training on mental abilities over fairly long periods of time. In these studies one is again faced with the lack of accurate instruments to measure mental abilities at various ages and a lack of agreement in the definitions of mental ability. Much of the present uncertainty of results will only be cleared up by studies which have completely fulfilled the prerequisites listed in Chapter 21 for the measurement of growth. Furthermore, nearly everyone who is familiar with mental testing knows that special coaching and practice increase the scores, both for tests of information and of comparison and reasoning skills, Chapter 20.

The number of studies which has been reported is not large, and the small number of persons usually appraised renders the studies inconclusive in many ways.

A good illustration of the effects of lack of school training is furnished by Gordon's (1923) report. Using the Stanford-Binet 1916 test, he found that the mean IQ's of children who lived on canal boats steadily decreased from 90 at ages 4 to 6, to 60 at ages 12 to 20. Gordon believed that the older children were not innately inferior to the younger, but that they had grown up in an environment where formal schooling was very rare and home life afforded very little intellectual training. Many of the parents of these children were nearly illiterate. Similar results were found for children of gypsies and for children with physical defects which prevented formal school training.

The effects of deafness upon mental test scores has been extensively reported by Pintner (1931). Of course, deaf children have a severe handicap in learning language, but they also were found to show great retardation in nonlanguage tests as well. A nation-wide survey under Pintner's direction applied the Pintner Nonlanguage Test to 4,432 deaf children, ages 12 or above, in 41 schools, and to a similar group of hearing children. All of the deaf age groups had mean IQ's of approximately 80, and mean EQ's of 62.[2]

The causes of these differences are not clear. It may be that a tendency to become deaf is associated with mental inferiority through the germ plasm. Congenital deafness is slightly more frequent in families with low mental ratings. It is also believed that about 85 per cent of deafness occurs after birth and that cerebral meningitis and scarlet fever may be important causes of both deafness and mental retardation. Lastly, it is probable that loss of hearing may arrest normal mental development through lack of normal stimulation.

The effects of particular sorts of schooling have been studied by a number of persons using in most instances IQ's from individual intelligence tests. Hildreth (1928) compared 48 pupils entering the 1st grade with no previous schooling with 41 pupils entering with 4 or 5 months of kindergarten or preschool experience. Both groups averaged six years of age, and both had unusually high social and economic home ratings. The group with no previous school experience had a mean IQ of 114 on entering the first grade, and the other group, 119. Eighteen months later the first group mean was 117 and the second, 119. The mean increase of 3 points in the first group during the eighteen-month period was not considered to be significant, and no significant differences were found between the two groups.

The influences of American rural and city environments upon Stanford-Binet scores have been compared by Jones. Conrad, and Blanchard (1932). They found that the rural children were relatively more backward at ages 5. 6. 7. and after age 11, than at ages 4, 8, 9, 10, and 11. Their tentative explanation was in part that at 4 years of age both rural and city children were usually out of school. The 5-, 6-, and 7-year-olds often go to school in the city but seldom in the poorer rural areas; hence the rural children fall behind in IQ's. During the next four years both groups are in school and the rural children become less retarded. Thereafter, school attendance in rural communities tends to be irregular and the quality of instruction poor. These investigators also point out that tests involving pencils, money, days of the week, dates, and a literary vocabulary introduce specific environmental aspects unfamiliar to many rural children.

---

[2] Upshall (1920) and Drever (1929) have reported that the deaf in day schools are superior intellectually to those in residential schools. even when equated for length of deafness.

Lamson (1938) found no gains in IQ's among children attending Grade IV of a demonstration school which had a greatly enriched curriculum. Peterson (1938) compared 20 children who had attended the nursery school at the University of Iowa with 15 children who had acquired no nursery-school experience at the time that both groups enrolled in the first grade. The groups were quite similar in social and economic backgrounds and in tests of reading readiness, vocabulary, and information. The children who had gone to nursery school were rated as more aggressive, sociable, independent, and socially competent than those who had not attended nursery school.

Several studies by Wellman have shown fairly large and permanent gains in IQ's which are apparently due to special schooling. In 1937 she reported the records of 78 persons who had been studied at frequent intervals from 5 to 18 years of age. Twenty-one of these had attended the Iowa Child Welfare Research Station Preschool and when tested at a mean age of 51 months had a mean IQ of 126. The other 53 students had not attended nursery school. Their first test, which yielded a mean IQ of 115, had been administered at a mean age of 72 months. The tests were either the 1916 Stanford Revision of the Binet, or, for children under 3½ years, the Kuhlmann-Binet. Both groups spent an average of seven years in the Iowa University School system, out of a possible fifteen years. The correlations between initial tests and centiles in the Iowa University College Entrance

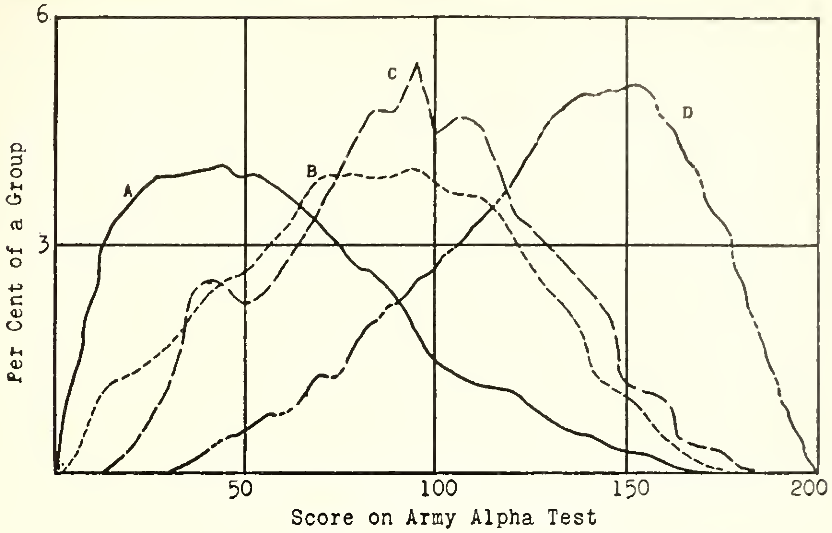
Examinations were .07 for those who attended preschool and .48 for those who did not have preschool experience. Forty-four of these two groups were retested with the Stanford Binet between 10 and 14 years of age. These IQ's correlated .56 with College Entrance Examination centiles. The years of attendance at the University schools correlated .40 with college centiles for the preschool group and .34 for the non-preschool group. The initial IQ's correlated –.07 with years of attendance. Another group of 82 children, divided into two similar groups, showed very similar correlations between IQ's and scores at the twelfth grade on the *American Council on Education Psychological Examination for College Freshmen.*

The mean scores indicated that attendance at any of the University schools tended to raise IQ's, and that, in general, the longer the attendance the greater the rise. Students who had attended the nursery school showed a consistent superiority at entrance into college over a non-nursery school group, when both groups were equal in initial IQ's and in attendance at University schools after the nursery school. The improvement in scores was larger among those with average IQ's. No inferior persons were included in the study and the superior persons had little scope for relative improvement as measured by the tests used. Wellman pointed out that although the number of cases studied was small, the results were unusually consistent in showing marked changes in IQ's which corresponded to superior educational environment. There were also marked improvements in interests, outlook on life, and other personal characteristics which showed themselves in behavior outside of formal test situations.

A study of the effects of practice upon the similarity of twins was reported by McNemar (1933). He set up an experiment to discover whether or not equal amounts of practice would increase or decrease the differences between members of twin pairs. He located 46 fraternal pairs and

47 identical pairs in seven California cities, and secured their co-operation in a seven-hour testing program which was administered during a single day. During this period a 20-second pursuit rotor test was given 70 times; a 60-second spool packing test, 28 times; and a card sorting test, 28 times. All of these tests showed marked improvement, particularly during the earlier trials. The twin resemblances as shown by correlations remained nearly the same for the identical twins on all tests; i.e., approximately .869 for the pursuit rotor test, .542 for spool packing, and .707 for card sorting. Among the fraternal twins resemblance correlations increased on the pursuit rotor test from .450 initially, to .614 finally. On the spool packing there was a slight increase in resemblance, from .413 to .592, and on the card sorting there was a slight decrease in resemblance, from .612 to .538. These changes in resemblance came rather early in the practice series. McNemar concluded that the changes noted in fraternal twin resemblances were too small to be very significant. The results, as far as they go, support the hypothesis that the environments of a pair of fraternal twins were not quite so alike as the environments of a pair of identical twins. On the other hand, since practice did not raise fraternal resemblance to that of the identical twins, he inferred that the differences in resemblances were largely hereditary.

Another set of figures from the United States Army Tests of 1917 and 1918 is often given as evidence that superior native ability will show itself in test scores in spite of educational handicaps. Some of the data, given in Figure 108, show that a large group of commissioned officers with less than ninth-grade education made nearly the same scores as a large group of enlisted men who had had formal education beyond the eighth grade. These figures are hard to evaluate because native mental ability is nowhere accurately defined. Moreover, the United States Army Alpha Test was designed to sample every-day adult experience as well as



	Median	Mean	SD	Number
A. Unselected, White, Native	58.9	65.2	38.2	51620
B. White draft with 9th-grade or more education	97.4	96.6	29.2	13943
C. Officers with less than 9th-grade education	107.3	104.3	30.1	660
D. All officers, unselected	139.3	132.5	26.7	12396

FIGURE 108. **Distribution of Army Alpha Scores**

Tables from Yerkes, ed., 1925, Part III, Chapter 10, have been combined

(By permission of the Editor, *Memoirs of the National Academy of Science*.)

formal school achievement. Hence a high test score may be indication of a stimulating or exacting job in an environment where persons are fairly well informed, rather than a superior native ability. There was a correlation of .75 between total Alpha scores and reported schooling in a native-born white sample of 653 enlisted men, but this figure is of little significance without a careful appraisal of the individual opportunities and motives for schooling and types of school attended.

BIBLIOGRAPHY

BIBLIOGRAPHY

Indebtedness to the authors and publishers of the following works is gratefully acknowledged. Wherever substantial quotations have been given, page references appear in the appropriate places in this text.

The page numbers for this text are printed at the end of each reference in italics. They refer only to the beginning of the discussion of the author's work although in some cases this discussion extends for several pages.

FREQUENT ABBREVIATIONS

Abn.	Abnormal	Med.	Medical
Acad.	Academy	Ment.	Mental
Acoust.	Acoustical	Minn.	Minnesota
Am.	American	Mo.	Monthly
Anthro.	Anthropological	Monog.	Monograph
Appl.	Applied	Nat.	National
Arch.	Archives	Nerv.	Nervous
Assn.	Association	Neur.	Neurology
B.	Bulletin	Ped.	Pedagogical
Brit.	British	Pers.	Personnel
Cath.	Catholic	Philos.	Philosophy
Clin.	Clinic	Physio.	Physiological
Col.	College	Psychiat.	Psychiatry
Comp.	Comparative	Psychol.	Psychology *or* Psychological
Contrib.	Contribution		
Corp.	Corporation	Pub.	Publication *or* Publishing
Dept.	Department		
Devel.	Development	Q.	Quarterly
Ed.	Education	Rec.	Record
Elem.	Elementary	Res.	Research
Empl.	Employment	Rev.	Review
Engl.	English	Sch.	School
Exp.	Experimental	Sci.	Science
Gen.	General	Sem.	Seminary
Genet.	Genetics	Soc.	Society
Hered.	Heredity	Stud.	Studies
Inst.	Institute	Suppl.	Supplement
J.	Journal	Teach.	Teachers
Juv.	Juvenile	Univ.	University
		Yr.	Year

Abernethy, E. M. (1936) Relationship between mental and physical growth. Soc. Res. Child Devel. Monog., 1, 7, 1–80. *660, 664*

Adams, H. F. (1927) The good judge of personality. J. Abn. and Social Psychol., 22, 172–181. *711*

——. (1930) An objectivity-subjectivity ratio for scales of measurement. J. Social Psychol., 1, 122–135. *706*

——. (1936) Validity, reliability, and objectivity. Psychol. Monog., 47, 329–350. *98, 105, 709, 711*

Adkins, D. C. (1937) The effects of practice on intelligence test scores. J. Ed. Psychol., 28, 222–231. *612*

Alexander, W. P. (1935) Intelligence, concrete and abstract. Brit. J. Psychol. Monog. Suppl., No. 19. *338*

Allison, A. (1821) Essays on the nature and principles of taste. Psychol. Monog., 51, 1–22. *387*

Allport, G. W. (1928) A test for ascendance-submission. J. Abn. and Social Psychol., 23, 118–136. *541, 571*

——. (1937) Personality: a psychological interpretation. Henry Holt, N. Y. Pp. 588. *520*

Allport, G. W., and Vernon, P. E. (1930) A study of values (test form). Houghton Mifflin, Boston. *486*

——. (1931) A test for personal values. J. of Abn. and Social Psychol., 26, 231–248. *502, 503*

American Council on Education. (1929) Civics and government test by Leigh, R. D., McGoldrick, J. D., Odegard, P. H., and Wood, B. D. World Book Co., Yonkers, N. Y., *238*

Anastasi, A. (1934) Practice and variability: a study in psychological method. Psychol. Monog., 45, 55. *615*

——. (1934 a) The influence of practice upon the test reliability. J. Ed. Psychol., 25, 321–335. *616*

——. (1937) Differential psychology. Macmillan, N. Y. Pp. 614. *623*

Anastasi, A., and Foley, J. P. (1936) Analysis of spontaneous drawings by children of different cultures. J. Appl. Psychol., 20, 689–726. *405*

Anderson, J. E. (1935) The effect of item analysis upon the discriminative power of an examination. J. Appl. Psychol., 19, 237–244. *102, 121*

Anderson, W. A. (1934) Occupational attitudes of college men. J. Social Psychol., 5, 435–465. *448, 450*

Andrew, D. M., and Bird, C. (1938) A comparison of two new type questions, recall and recognition. J. Ed. Psychol., 29, 175–193. *129*

Andrew, D. M., Paterson, D. G., and Longstaff, H. P. (1933) Minnesota vocational test for clerical workers. The Psychol. Corp., N. Y., *23*

Arrington, Ruth E. (1932) Interrelations in the behavior of young children. Teach. Col., Columbia Univ., Child Devel. Monog., No. 8, 20, 156. *176, 177*

Arthur, Grace. (1930) A point scale of performance tests. Commonwealth Fund. Pp. 82. *673*

Asch, S. E. (1936) A study of change in mental organization. Arch. Psychol., No. 195. Pp. 30. *253, 255, 663*

Austin, S. B. McK. (1921) A study of logical memory. Am. J. Psychol., 32, 270–403. *634*

Aveling, Francis. (1926) Psychology of conation and volition. Brit. J. Psychol., 16, 339–353. *591*

Ayer, E. C. (1916) Psychology of drawing. Warwick York, Baltimore. Pp. 186. *405*

Ayres, G. H. (1934) Predicting success in algebra. Sch. and Soc., 39, 17–18. *257*

Ayres, L. P. (1912) A scale for measuring the quality of handwriting of school children. Russell Sage Found., N. Y. Pp. 90. *194, 195*

——. (1915) A measuring scale for ability in spelling. Russell Sage Found., N. Y. Pp. 58. *216, 676*

Bachern, A. (1937) Various types of absolute pitch. J. Acoust. Soc. Am., 9, 146–151. *427*

Baker, H. J., and Crockett, A. C. (1928) Detroit mechanical aptitudes examination for boys. Public Sch. Pub. Co., Bloomington, Ill. *30, 356, 376*

Baker, H. J., and Leland, B. (1935 a) Detroit learning aptitude tests. Public Sch. Pub. Co., Bloomington, Ill. Pp. 157. *24, 275, 276, 282, 291*

Baker, H. J., and Traphagen, Virginia. (1935) Diagnosis and treatment of behavior-problem children. Macmillan Co., N. Y. Pp. 393. *566*

Baker, H. J., and Voelker, P. H. (1937) Detroit mechanical aptitudes examination. Public Sch. Pub. Co., Bloomington, Ill. *356*

Baldwin, B. T., and Stecher, L. I. (1924) The psychology of the preschool child. Appleton-Century, N. Y. Pp. 296. *151*

Barlow, M. C. (1924) Individual differences as affected by continued practice. Master's thesis. George Peabody College for Teachers. *623*

Barrett, E. R., Ryan, T. M., and Schrammel, H. E. (1938) Barrett-Ryan-Schrammel English test. World Book Co., Yonkers, N. Y. *211*

Basedow, J. B. (1770) A methodenbuch für vater und mutter der familien und volker. Ansgewählte Scriften, Bibliotek Pädagogischer Klassiker, Langensala. *151*

Bayley, Nancy (1933) Mental growth during the first three years. A developmental study of 61 children by repeated tests. Genet. Psychol. Monog., 14, No. 1. Pp. 92. *182, 657*

——. (1935) Some comparisons between growth in motor and mental abilities in young children. Psychol. B., 31, 608. *369*

——. (1940) Mental growth in young children. 39th Yr. Book, Nat. Soc. Study Ed., Part II, 11–13. *183, 184*

Bean, K. L. (1936) Musical talent of the southern negroes as measured by the Seashore tests. J. Genet. Psychol., 49, 244–249. *437*

Beck, S. J. (1932) The Rorschach test as applied to a feeble-minded group. Arch. Psychol., 136. Pp. 84. *524, 529, 531, 532*

——. (1937) Introduction to the Rorschach method. Monog. No. 1, The Am. Orthopsychiatric Assn. *525, 528*

Beebe-Center, J. G., and Pratt, C. C. (1937) A test of Birkoff's aesthetic measures. J. Genet. Psychol., 17, 339–353. *432*

Beers, F. A., and Paterson, D. G. (1933) Cooperative literary acquaintance test. Cooperative Test Service, N. Y. *423*

Bell, H. M. (1934) Adjustment inventory. Stanford Univ. Press. *545, 546*

Benton, A. L. (1935) The interpretation of questionnaire items in a personality schedule. Arch. Psychol., 28, No. 190, 38. *595*

Berger, H. (1937) Über das Elektrent-Kephalogramm des Menschen. Arch. Psychiat. Nervenkrankh, 106, 13, 577–584. *367*

Berliner, Anna. (1918) Aesthetic judgments of school children. J. Appl. Psychol., 2, 229–242. *396*

Bernreuter, R. G. (1931) Personality inventory. Stanford Univ. Press. *541, 591*

Bernreuter, R. G., and Carr, E. J. (1938) The interpretation of IQ's and the L–M Stanford-Binet. J. Ed. Psychol., 29, 312–314. *290*

Berry, C. S. (1923) The intelligence quotients of mentally retarded children. Sch. and Soc., 17, 723–729. *294*

Betts, E. A. (1936) Ready to read tests. Keystone View Co., Meadville, Pa. *175*

Biddle, W. W. (1931) The relationship between knowledge and a measure of autistic thinking on certain international problems. J. Social Psychol., 2, 193–196. *515*

——. (1932) Propaganda and education. Contrib. to Ed., Teach. Col., Columbia Univ. Press, No. 531. *511*

Bills, A. G. (1937) Fatigue in mental work. Physiol. Rev., 17, 436–453. *376*

Binet, A. (1898) La mésure en psychologie individuelle. Revue Philosophique, 46, 2d Sem. 113–123. *265*

——. (1900) Attention et adaptation. L'Année Psychologique, 6, 248–404. *265*

Bingham, W. V. (1937) Aptitudes and aptitude testing. Harper, N. Y. Pp. 390. *93*

Bird, M. H. (1932) A study in aesthetics. Harvard Univ. Press, Cambridge, Mass. Pp. 117. *396, 409, 410*

Birkhoff, G. D. (1933) Aesthetic measure. Harvard Univ. Press, Cambridge, Mass. Pp. 226. *432*

Bisbee, E. U. (1933) Commercial education survey tests: Junior and senior shorthand. Public Sch. Pub. Co., Bloomington, Ill. *242*

Blackstone, E. G. (1923) Blackstone stenographic proficiency tests: Typewriting test. World Book Co., Yonkers, N. Y. *242*

Blackstone, E. G., and McLaughlin, M. W. (1932) Blackstone stenographic proficiency tests: Stenography test. World Book Co., Yonkers, N. Y. *242*

Bleuler, M., and Bleuler, R. (1935) Rorschach ink-blot test and racial psychology. Character and Personality, 4, 97–114. *530*

Bobertag, O. (1912) Über Intelligenzprufungen nach der Methode von Binet und Simon. Zeitschrift fur angewandte Psychologie, 5, 495–538. *268*

Bolton, E. B. (1935) Effect of knowledge upon attitudes toward the negro. J. Social Psychol., 6, 68–90. *515*

Bovard, J. F., and Cozens, F. W. (1938) Tests and measurements in physical education. Saunders. Pp. 427. *371*

Bradshaw, F. F. (1930) American council on education rating scale: its reliability, validity, and use. Arch. Psychol., No. 119. Columbia Univ., New York. *706*

Brandt, H. F. (1937) A logical decision test: an indirect approach to the study of the emotional orientation of the adolescent. University of Iowa Study in Child Welfare Studies in Emotional Adjustment. 13, No. 4, 9–24. *486, 498, 500*

Bregman, E. O. (1935) Revision of the army alpha examination, form B. Psychol. Corp., N. Y. *319*

Breidenbaugh, U. E. (1940) Breidenbaugh bookkeeping tests. Public Sch. Pub. Co., Bloomington, Ill. *244*

Brennan, F. M. (1926) The relation between musical capacity and performance. Psychol. Monog., 36, No. 167, 190–248. *427*

Brigham, C. C. (1923) A study of American intelligence. Princeton Univ. Press. Pp. 210. *319*

Brighouse, G. (1939) Variability in preferences for simple forms. Psychol. Monog., 51, 68. *390*

Brogden, H. E. (1940) A factor analysis of forty character tests. Psychol. Monog., 52, No. 3, 39–56. *588*

Brooks, F. D. (1929) The psychology of adolescence. Houghton Mifflin, Boston, 6, 652. *611*

——. (1937) Child psychology. Houghton Mifflin, Boston. Pp. 599. *153, 663*

Brown, A. W. (1936) Chicago non-verbal examination. Inst. Juvenile Res., Chicago. *352*

Brown, A. W., and Woody, C. (1928) Brown-Woody civics test. World Book Co., Yonkers, N. Y. *238*

Brown, W. (1910) Some experimental results in the correlation of mental abilities. Brit. J. Psychol., 3, 296–322. *298*

Brown, Wm., and Thomson, G. H. (1921) The essentials of mental measurement. Univ. Press, Cambridge, England. Pp. 216. *105*

Bryngelson, B. (1932) A photophonographic analysis of the vocal disturbances in stuttering. Psychol. Monog., 43, No. 194, 1–30. *193*

Buck, W. (1936) A measurement of changes in attitudes and interests of University students over a ten-year period. J. Abn. and Social Psychol., 31, 12–19. *475*

Buckingham, B. R. (1927) Extension of the Ayres spelling scale. Public Sch. Pub. Co., Bloomington, Ill. *217*

Buckingham, G. E. (1936) Diagnostic and remedial teaching in first year algebra. J. Ed. Res., 30, 198–213. *253*

Buckingham, G. E., and Dolch, E. W. (1936) Combined word list. Ginn, Boston. Pp. 185. *200*

Buckingham, G. E., and Lee, R. E. (1936) A technique for testing unified concepts in science. J. Ed. Res., 30, 20–27. *234*

Bühler, C., and Hetzer, H. (1935) Testing children's development from birth to school age. Farrar and Rinehart, N. Y. Pp. 191. *158, 160, 161, 182*

Buhler, C., and Others. (1939) The child and his family. Harper, N. Y. Pp. viii + 187. *562*

Bullough, E. (1908) The 'perceptive problem' in the aesthetic appreciation of single colours. Brit. J. Psychol., 2. 406–462. *698*

——. (1910) The perceptive problem in the aesthetic appreciation of simple color combinations. Brit. J. Psychol., III, 406–447. *420, 429*

Burks, B. (1928) The relative influence of nature and nurture upon mental development. 27th Yr. Book, Nat. Soc. Study Ed., Part 1, 219–315. *716, 719*

Burnham, P. S., and Crawford, A. B. (1935) The vocational interests and personality test scores of a pair of dice. J. Ed. Psychol., 26, 508–512. *581*

Burns. Z. H. (1937) Practice, variability and motivation. J. Ed. Res., 30, 403–420. *624*

Buros, O. K. (1938) Mental measurements yearbook. Issued yearly. Rutgers Univ. Press. Pp. 115. *153*

Burt, C. (1909) Experimental tests of general intelligence. Brit. J. Psychol. 3, 94–177. *299*

——. (1922) Mental and scholastic tests. P. S. King and Son, Ltd., London. Pp. 432. *277, 280, 287*

——. (1938) Analysis of temperament. Brit. J. Med. Psychol., 17, 158–188. *141, 588*

——. (1939) The relation of educational abilities. J. Ed. Psychol., 9. 45–71. *254*

Buswell, G. T. (1922) Fundamental reading habits: a study of their development. Suppl. Ed. Monog., No. 21, Univ. of Chicago Press, 149. *655*

Buswell, G. T., and John, L. (1925) Diagnostic chart for individual difficulties fundamental processes in arithmetic. Public Sch. Pub. Co., Bloomington, Ill. *229, 230*

Buxton, C. (1938) The application of multiple factorial methods to the study of motor abilities. Psychometrika, 3, 85–93. *380*

Cahalan, E. J. (1939) The consistency of aesthetic judgment. Psychol. Monog., 51, 1–75. *391*

Cannon, A. (1937) Testing the appreciation of style. Engl. J., 26, 648–654. *418*

Cantril, H., and Allport, G. W. (1933) Recent applications of the study of values. J. Abn. and Soc. Psychol., 28, 259–273. *504*

Cantril, H., and Rand, H. G. (1934) An additional study of the determination of personal interests by psychological and graphological methods. Char. and Pers., 3, 72–78. *193*

Carlson, H. B. (1934) Attitudes of undergraduate students. J. Social Psychol., 5, 202–213. *509*

Carmichael, L. (1936) A re-evaluation of the concepts of maturation and learning as applied to the early development of behavior. Psychol. Rev., 43, 450–470. *713*

Carroll, H. A. (1935) Prose appreciation test. Test Bureau, Minneapolis. *421, 422*

Carroll, H. A., and Eurich, A. C. (1932) Intelligence and art appreciation. J. Ed. Psychol., 23, 214–221. *394*

Carter, H. D., Conrad, H. S., and Jones, M. C. (1935) A multiple factor analysis of children's annoyances. J. Genet. Psychol., 47, 282–298. *588*

Carter, H. D., Pyles, M. K., and Bretnall, E. P. (1935) A comparative study of factors in vocational interest scores of high school boys. J. Ed. Psychol., 26, 81–98. *471*

Cattell, J. McK. (1902) The time of perception as a measure of differences in intensity. Philos. Stud., 19, 63–68. *692*

——. (1903) A statistical study of eminent men. Pop. Sci. Monthly, 62, 359–377. *694, 714*

Cattell, J. McK., and Bryant, Sophie. (1889) Mental association investigated by experiment. Mind, 14, 230–250. *298*

Cattell, P. (1928) Dentition as a measure of maturity. Harvard Monog. Ed., No. 9, Harvard Univ. Press, Cambridge, Mass., 91. *646*

——. (1931) Constant changes in the Stanford-Binet I.Q. J. Ed. Psychol., 22, 544–550. *293, 294*

Cattell, R. B. (1934) Temperament tests. II Tests. Brit. J. Psychol., 24, 20–49. *589*

Cattell, R. B. (1936a) Is national intelligence declining? J. Hered., 27, 181–203. *718*

——. (1936) Temperament tests in clinical practice. Brit. J. Med. Psychol., 16, 43–61. *540*

Champney, H., and Marshall, H. (1939) Optimal refinement of the rating scale. J. Appl. Psychol., 2, 323–331. *704*

Chandler, A. R. (1934) Beauty and human nature: elements of psychological aesthetics. Psychol. Monog., 51, 22. *385, 387*

Chapman, J. C. (1914) Individual differences in ability and improvement and their correlations. Contrib. to Ed. Teach. Col., Columbia Univ., No. 63. *623*

——. (1921) Trade tests. Henry Holt, N. Y. *317*

——. (1924) Chapman-Cook speed of reading test. Grades IV to VIII. Lippincott. Phila. *224*

Charters, W. W. (1919) Errors in language. Pittsburgh Schools. Unpublished report on file with Board of Education, Pittsburgh. *210*

Chase, W. P. (1937) Color vision in infants. J. Exp. Psychol., 20, 203–222. *401*

Cherrington, B. M., and Miller, L. W. (1933) Changes in attitude as the result of a lecture and reading similar materials. J. Social Psychol., 4, 479–484. *512*

Christensen, E. O., and Karwoski, T. (1925) A test in art appreciation. Art Psychol. B., 9, 1 and 3. Univ. N. Dakota. *395*

Clair, M. B. (1939) Variation in the perception of aesthetic qualities in paintings. Psychol. Monog., 51, 52. *391, 392*

Clapp, F. L., and Young, P. (1929) Self marking tests. Houghton Mifflin, Boston. *132*

Clark, W. H. (1935) Two tests for perseverance. J. Ed. Psychol., 26, 604–610. *589*

Cleeton, G. U. (1937) Vocational interest inventory. McKnight and McKnight, Bloomington, Ill. *467, 469*

Coghill, G. E. (1926) Correlated anatomical and physiological studies of the growth of the nervous system in Amphibia. VI. The mechanism of integration in Amblystoma punctatum. J. Comp. Neur., No. 1. *639*

——. (1930) Individuation vs. integration in the development of behavior. J. Gen. Psychol., 3, 431–435. *639*

Cohn, J. (1894) Experimentelle Untersuchungen über die Gefuhlsbetonung der farben, helligkeiten und ihrer combinationen. Phil. Stud., 10, 562–603. *692*

Columba, M. (1926) A study of interests and their relations to other factors of achievement in the elementary school subjects. Cath. Univ. Am., Ed. Res. B., 1, No. 7, 35. *476, 478*

Comenius, J. A. (1628 reprint 1896) Comenius school of infancy. Heath, Boston. Pp. 99. *151*

Conklin, E. G. (1922) Heredity and environment. Fifth Edition. Princeton Univ. Press. *638*

Conklin, E. S., and Sutherland, J. W. (1923) A comparison of the scale of values method with the order of merit method. J. Exp. Psychol., 6, 44–57. *698*

Conrad, H. S. (1932) The personal equation in ratings: I. An experimental determination. Ped. Sem. and J. Genet. Psychol., 41, 267–293. *708*

——. (1932 a) The validity of personality ratings of nursery-school children. J. Ed. Psychol., 23, 671–680. *708*

Courtis, S. A. (1925) Why children succeed. Courtis standard tests, Detroit. Pp. 271. *591*

——. (1929) Maturation units for the measurement of growth. Sch. and Soc., 30, 683–690. *668*

——. (1932) Measurement of growth. Brumfield and Brumfield, Ann Arbor. Pp. 221. *668, 670*

——. (1933) The prediction of growth. J. Ed. Res., 26, 481–492. *608, 611, 629, 669*

——. (1935) Maturation as a factor in diagnosis. Yearbook Nat. Soc. Study Ed., 34, 169–187. *634*

——. (1937) Measurements of dentition. Pp. 34. Manuscript in author's file. *647*

——. (1938) Maturation units for the measurement of growths. Sch. and Soc., 30, 683–690. *649, 656, 671, 672*

Cowdery, K. M. (1926) Measurement of professional attitudes. Differences between lawyers, physicians, engineers. J. Per. Res., 5, 131–141. *480*

Crissey, O. L. (1937) Mental development as related to institutional residence and educational achievement. Univ. of Iowa, Stud. Child Welfare, 13, 1, 1–81. *721*

Crissey, W. J. E., and Daniel, W. J. (1939) Vocational interest factors in women. J. Appl. Psychol., 23, 488–494. *472*

Curtis, F. D., and Woods, G. G. (1928) A study of a modified form of the multiple-response test. J. Ed. Res., 18, 211–219. *114.*

Daniels, P. C. (1933) Discrimination of compositional balance at the preschool level. Psychol. Monog., 1. *405*

Davenport, C. B. (1912) Heredity in relation to eugenics. Henry Holt, N. Y. Pp. 298. *715*

Davis, H. (1938) Interpretation of the electrical activity of the brain. Am. J. Psychiat., 94, 825–834. *367*

Dawson, H. L. (1936) Technics for measuring infants. J. Pediatrics, 9, 187–202. *159*

Dearborn, W. F., and Anderson, J. H. (1938) Controlled reading by means of a motion picture technique. Psychol. Rec. 11, 219–227. *226*

Dearborn, W. F., Anderson, J. E., and Christiansen, A. O. (1916) Construction tests of mental ability. J. Ed. Psychol., 7, 115–158. *341*

de Sanctis, Sante. (1906) Types et degrés d'insuffisance mentale. L'Année Psychologique, 12, 70–83. *268*

Descoeudres, A. (1911) Les Tests de Binet and Simon et leur valeur scolaire. Archives de Psychologie, II, 331–350. *268*

Dewey, E. (1935) Behavior development in infants. Columbia Univ. Press. Pp. 321. *153*

Dexter, E. S. (1932) On resemblance between mother-daughter and sister-sister pairs. Sch. and Soc., 35, 906, 640. *716*

Dodd, S. C. (1926) International group mental tests. Princeton, N. J. Miméographed only. *352*

Downey, June E. (1927) Individual differences in reaction to the word-in-itself. Am. J. Psychol., 39, 323–342. *420*

Dreps, H. F. (1933) The psychological capacities and abilities of college art students of high and low standing. Psychol. Monog., 45, 1, 134–146. *412*

Drever, J. (1929) Intelligence tests for the deaf. Teach. of the Deaf, 27, 163–167. *725*

Drought, R. Alice. (1929) A survey of studies in experimental aesthetics. J. Ed. Res., 20, 97. *395*

Duff, L. F. G. (1937) A one-sided sketch of Jonathan Swift. Psychoanalytic Q., 6, 238–259. *425*

Dunlap, J. W. (1938) Relationship between type of question and scoring errors. J. Exp. Ed., 6, 376–379. *111*

Durrell, D. D. (1935) Tests and corrective procedures for reading disabilities. Elem. Engl. Rev., 12, 4, 91–95. *193*

Dvorak, Beatrice J. (1935) Differential occupational ability patterns. B. Empl. Stabilization Res. Inst., Vol. 3, No. 8, Univ. of Minn. Press. Pp. 47. *62*

Dwyer, P. S. (1938) An analysis of 19 occupational scores of the Strong vocational interest test given to 118 students entering the University of Michigan medical school during the years 1928, 1929, 1930. J. Appl. Psychol., 22, 8–16. *472*

East, E. M. (1936) Genetic aspects of certain problems of evolution. Am. Nat., 70, 143–158. *713*

Ebbinghaus, H. (1885) Über die Gedachtniss. Trans. 1913 by H. A. Ruger. Teach. Col., Columbia Univ., New York. *634*

——. (1897) Uber eine neue methude zur prüfung geistigen fahigkeiten und ihre anwendungbei schulkindern. Zeitschrift für angewandte Psychologie, 13, 401–459. *298*

Ebert, E., and Meumann, E. (1905) Ueber einige Grundfragen der Psych. der Uebungsphanomene im Bereiche des Gedachtnisses. Arch. f. Gesante Psychologie, 4, 232. *298*

Edmonds, E. M., and Smith, M. E. (1923) The phenomenological description of musical intervals. Amer. J. Psychol., 34, 287–291. *429*

Elwell, F. H., and Fowlkes, J. G. (1928) Bookkeeping test 2A. World Book Co., N. Y. *245*

Estabrook, A. H. (1916) The Jukes in 1915. Carnegie Inst., Washington, D. C. Pp. 85. *715*

Estes, S. G., and Horn, D. (1938) Interest patterns as related to fields of concentration among engineering students. J. of Psychol. (1939), 7, 29–36. *465*

Eurich, A. C. (1931) Reading abilities of college students. Univ. Minn. Press. Pp. 208. *224*

Exner, S. (1868) Ueber die zu einer gesichtswahrehmung nothige Zeit. Sitz. Ber. d. Wiener Akad., 58 (2), 601–632. *366*

Farmer, E. A. (1930) Note on the relation of certain aspects of character to industrial proficiency. Brit. J. Psychol., 21, 46–49. *362*

Farnsworth, P. R. (1931) An historical, critical, and experimental study of the Seashore-Kwalwasser test battery. Genet. Psychol. Monog., 9, 5, 291–389. *427, 433, 434*

——. (1938) Genetic study of the Bernreuter personality inventory. J. Genet. Psychol., 52, 3–13. *542*

Fechner, G. T. (1871) Zur experimentalen Aesthetik, Abhandl. math. phys. Classe d. K. Sächs. Ges. d. Wiss., 9, 555–635. *692*

Ferguson, G. O. (1920) A series of form boards. J. Exp. Psychol., 3, 47–58. *344*

Ferguson, L. W. (1939) The requirements of an adequate attitude scale. Psychol. B. 36, 665–673. *698*

Ferrari, G. C. (1908) Come si mesura lo sviluppo dell' intelligenza nei bambini normali. Rivista de psicologia, 4, 465–471. *268*

Finch, F. H., and Nemzek, C. L. (1934) Prediction of college achievement from data collected during the secondary school period. J. of Appl. Psychol., 18, 454–460. *257*

Flanagan, J. C. (1935) Factor analysis in the study of personality. Stanford Univ. Press. Pp. 103. *542*

——. (1939) Measuring interests. Psychol. B., 36, 529–530. *470*

Flory, C. D. (1936) Osseous development in the hand as an index of skeletal development. Soc. Res. Child Devel., 1, 3. Pp. 141. *645, 661*

Fox, C. (1938) The method of testing literary appreciation. Brit. J. Psychol., 29, 1–11. *423*

Franzen, R. H. (1922) The accomplishment ratio. Contrib. to Ed., Teach. Col., Columbia Univ., No. 125. *250*

Freeman, F. N. (1914) Freeman chart for diagnosing faults in handwriting. Range: all grades. Houghton Mifflin, Boston. *194*

——. (1916) Experimental education. Houghton Mifflin, Boston. Pp. 220. *358*

Freeman, F. N., *et al.* (1928) The influence of environment on the intelligence, school achievement and conduct of foster children. 27th Yr. Book. Nat. Soc. Stud. Ed., Part 1, 103–218. *716, 719*

Freeman, F. N., and Flory, C. D. (1937) Growth in intellectual ability as measured by repeated tests. Soc. Res. Child Devel., Nat. Res. Council, Washington, D. C. *650*

Freud, Anna (1925) Introduction to the technique of child analysis. Nervous and Mental Diseases Monog., No. 48. Pp. 62. *151, 181*

Freyd, M. (1922) A method for the study of vocational interests. J. Appl. Psychol., 6, 243–254. *460*

Froebel, F. (1826) The education of man. Appleton-Century, N. Y. Pp. 332. *151*

Fryer, Douglas. (1927) Predicting abilities from interests. J. Appl. Psychol., 11, 212–225. *478, 480*

——. (1931) The measurement of interests. Henry Holt, N. Y. Pp. 448. *80, 442, 448, 476*

Furfey, P. H., and Muehlenbein, J. (1932) The validity of infant intelligence tests. J. Genet. Psychol., 40, 219–223. *183*

Galton, F. (1869, 2nd Ed. 1892) Hereditary genius. Macmillan, N. Y. Pp. 379. *523, 714*

——. (1886) The inheritance of stature. J. Anthro. Inst., London. Vol. 15. *65, 66, 67, 68*

Gardner, G. E. (1936) Rorschach test replies and results in 100 normal adults of average I.Q. Am. J. Orthopsychiatry, 6, 32–60. *528*

Garfiel, E. (1923) The measurement of motor ability. Arch. Psychol., Columbia Univ., No. 52. *375, 378, 379*

Garretson, O. K., and Symonds, P. M. (1930) Interest questionnaire for high-school students. Bur. of Pub., Teach. Col., Columbia Univ. *458, 459, 480*

Garrett, H. E. (1928) The relation of tests of memory and learning to each other and to general intelligence in a highly selected adult group. J. Ed. Psychol., 19, 601–613. *253*

——. (1937) (1939) Statistics in psychology and education. Revised edition. Longmans Green, N. Y. Pp. 493. *57, 79, 84, 85, 87, 101, 105*

——. (1938) Differentiable mental traits. Psychol. Record, 2, 259–298. *663*

Garrett, H. E., Bryan, A. I., and Perl, R. E. (1935) The age factor in mental organization. Arch. Psychol., No. 175, 1–29. *253*

Gates, A. I. (1922) A study of reading and spelling with special reference to disability. J. Ed. Res., 6, 12–24. *257*

——. (1922 a) The correlations of achievement in school subjects with intelligence tests and other variables. J. Ed. Psychol., 12, 303–314, 378–391, 445–464. *216, 623*

——. (1926) Gates silent reading test. Bur. Publications, Teach. Col., Columbia Univ. *221, 222*

——. (1928) The nature and limit of improvement due to training. Public Sch. Pub. Co., Bloomington, Ill., 441–460. *372, 615*

——. (1930) Interest and ability in reading. Macmillan, N. Y. Pp. 264. *446*

Gaw, E. A. (1928) Five studies of the music tests. Psychol. Monog., 39, 178, 145–156. *433*

Gesell, Arnold. (1925) The mental growth of the pre-school child. Macmillan, N. Y. Pp. 447. *159, 182*

——. (1928) Infancy and human growth. Macmillan, N. Y. Pp. 418. *274*

——. (1929) The early diagnosis of mental defect. Arch. Neurol. and Psychol., 22, 522–529. *185*

——. (1940) Stability of mental growth careers. Nat. Soc. Stud. Ed. Yr. Book, 39, 149–159. *185*

Gesell, A., Thompson, H., and Amatruda, C. S. (1938) Psychology of early growth. Macmillan, N. Y. Pp. 290. *153, 155, 157, 158, 159, 182*

Gibbs, F. A. (1937) Interpretation of the electroencephalogram. J. Psychol., 4, 365–382. *367*

Giles, G. R. (1936) A new interest test. J. Ed. Psychol., 27, 527–536. *533*

Givler, R. C. (1915) The psychophysiological effect of the elements of speech in relation to poetry. Psychol. Monog., 19, No. 2, 132. *420*

Goddard, H. H. (1910) A measuring scale for intelligence. The Train. Sch., Vineland, N. J., No. 6, 146–155. *152*

——. (1911) Two thousand normal children measured by the Binet measuring scale of intelligence. Ped. Seminary, 18, 232–259. *268*

——. (1912) The height and weight of feeble-minded children in American institutions. J. of Nerv. and Ment. Dis., 39, 217–235. *648*

——. (1913) Feeblemindedness, its causes and consequences. Macmillan, N. Y. Pp. 599. *268, 715*

Goodenough, Florence L. (1926) Measurement of intelligence by drawings. World Book Co., Yonkers, N. Y. Pp. 177. *167*

——. (1928) Measuring behavior traits by means of repeated short samples. J. Juv. Res., 12, 230–235. *168, 405, 408*

——. (1930) Inter-relationships in the behavior of young children. Child Devel., 1, 29–48. *569*

——. (1931) Chapter on children's drawings in Handbook of Child Psychology. Clark Univ. Press. Pp. 711. *410*

Goodenough, F. L., Foster, J. G., and Van Wagenen, M. J. (1932) The Minnesota pre-school tests. Ed. Test Bur., Minneapolis, Minn. Pp. 44. *165, 166, 167, 614*

Gordon, H. (1923) Mental and scholastic tests among retarded children. Board of Ed. Pamphlets, London, No. 44, 1–92. *724*

Gray, W. S. (1917) Studies of elementary school reading through standardized tests. Suppl. Ed. Monog., Vol. 1, No. 1, Univ. of Chicago. *193*

Greene, E. B. (1927) The effect of time limits upon U. S. army alpha score. Manuscript in author's files. *96*

——. (1929) Achievement and confidence on true-false tests of college students. J. of Abn. and Social Psychol., 23, 4, 467–478. *130*

——. (1931) Michigan nonverbal series. Edwards Bros., Ann Arbor, Mich. *18, 352, 353, 372, 689*

——. (1931 a) The retention of information learned in college courses. J. Ed. Res., 24, 262–273. *635, 636, 637*

——. (1932) Michigan nonverbal series. J. of Social Psychol., 3, 354–359. *683*

——. (1933) Effects of practice on growth curves. Manuscript in author's file. *612, 613*

——. (1934) Michigan speed of reading tests. J. Ed. Res., 28, 283–288. *654*

——. (1936) Practice and performance. Occupations, 15, 51–56. *605, 623*

——. (1937) Practice effects on various types of standard tests. Am. J. Psychol., 49, 67–75. *611, 617, 620, 622*

——. (1937 a) Analysis of adjustments to test situations. Manuscript in author's files. *576*

——. (1937 b) Michigan speed of reading tests, Forms 1 and 2. Psychol. Corp., N. Y. *48*

——. (1938) Michigan vocabulary profile. J. of Higher Ed., 9, 383–389. *130, 133, 449, 677*

——. (1938 a) Reasons for choice of vocation among college students. Pp. 16. Manuscript in author's files. *450*

——. (1939) The Michigan vocabulary profile tests. World Book Co., Yonkers, N. Y. Pp. 21. *19, 101, 206, 208, 210, 381, 181, 632*

Greene, H. A. (1918) A standardization of certain opposites tests. J. Ed. Psychol., 9, 559–566. *650*

Greene, H. A., and Jorgensen, A. N. (1927) Iowa silent reading test. Bur. Ed. Res. and Service, Univ. of Iowa. *206*

——. (1936) The use and interpretation of high school tests. Longmans, Green, N. Y. Pp. 614. *257*

Grippen, V. B. (1933) A study of creative artistic imagination in children by the constant contact procedure. Psychol. Monog., 45, 1, 63–81. *403, 408*

Guernsey, M. (1928) The role of consonance and dissonance in music. Am. J. Psychol., 40, 173–204. *428*

Guilford, J. P. (1936) Psychometric methods. McGraw Hill, N. Y. Pp. 565. *105, 138, 585, 588, 693, 695*

Guilford, J. P., and Guilford, R. B. (1931) A prognostic test for students in design. J. of Appl. Psychol., 25, 335. *397*

Guttman, E., and Maclay, W. S. (1937) Clinical observations on schizophrenic drawings. Brit. J. Med. Psychol., 16, 184–204. *533*

Haggerty, M. E. (1920) Haggerty intelligence examination. Delta 2. World Book Co., Yonkers, N. Y. *320*

Haggerty, M. E., Olson, W. C., and Wickman, E. K. (1930) Haggerty-Olson-Wickman behavior rating schedules. World Book Co., Yonkers, N. Y. *564, 565*

Halverson, H. M. (1931) An experimental study of prehension in infants. Genet. Psychol. Monog., 10, Nos. 2, 3. *177*

Hanks, L. M. (1936) Prediction from case material to personality test data: a methodological study of types. Arch. Psychol., No. 207. Pp. 71. *582*

Harrell, W. (1939) A factor analysis of mechanical ability tests. Psychol. B., 36, 521. *360, 382*

Harris, J. A., Jackson, C. M., Patterson, D. G., and Scammon, R. E. (1930) The measurement of man. Univ. of Minn. Press. Pp. 215. *644*

Hartshorne, H., and May, M. (1928) Studies in deceit. Macmillan Co., N. Y. Pp. 306. *549, 550, 552, 580*

——. (1929) Studies in service and self-control. Macmillan, N. Y. Pp. 559. *549, 554, 597*

——. (1930) III. Studies in the organization of character. Macmillan Co., N. Y. Pp. 503. *549, 556, 558, 560, 561, 580, 582, 591*

Harvey, O. L. (1934) Measurement of handwriting considered as a form of expressive movement. Character and Personality, 2, 310–321. *194*

Healy, W., and Fernald, G. (1911) Tests for practical mental classification. Psychol. Monog., No. 2. *344, 358*

Heinis, H. (1926) A personal constant. J. of Ed. Psychol., 17, 163–186. *294, 673*

Hendricks, H. B., Frutchey, F. P., Horney, A. G., and Tyler, R. W. (1934) Cooperative chemistry test. Co-operative Test Service, N. Y. *27*

Henmon, V. A. C., and Nelson, M. S. (1931) The Henmon-Nelson tests of mental ability. Houghton Mifflin Co., N. Y. *321*

Herbart, J. F. (1898) The application of psychology to the science of education. Scribner, N. Y., 231. *151*

Hertz, M. R. (1934) The reliability of the Rorschach ink blot test. J. Appl. Psychol., 18, 461–477. *527, 529*

——. (1935) Rorschach norms for an adolescent age group. Child Devel., 6, 69–76. *529*

——. (1936) The method of administration of the Rorschach ink-blot test. Child Devel., 7, 237–254. *528*

Hertz, M. R. (1938) Scoring the Rorschach test with specific reference to "Normal Detain" category. Am. J. of Orthopsychiatry, 8, 100–121. *528*

Hertzman, Max. (1936) The effects of the relative difficulty of mental tests on patterns of mental organization. Arch. Psychol., 28, 197, 69. *340, 633*

Hevner, K. (1930) An empirical study of three psychophysical methods. J. Gen. Psychol., 4, 191–212. *698*

Highsmith, J. A. (1929) Selecting musical talent. J. Appl. Psychol., 13, 486–493. *437*

Hildreth, G. (1925) Resemblance of siblings in intelligence and achievement. Contrib. to Ed., Teach. Col., Columbia Univ., No. 186. Pp. 65. *716*

——. (1928) The effect of school environment upon Stanford-Binet tests of young children. 27th Yr. Book, Nat. Soc. Stud. Ed., Part 1, 355–359. *725*

Hildreth, G. H., Griffiths, N. L., and Orleans, J. S. (1933) Metropolitan readiness test. World Book Co., Yonkers, N. Y. Pp. 16. *171*

Hillegas, M. B. (1912) A scale for the measurement of quality in English composition. Teach. Col. Rec., 13, No. 4. *195, 696*

Hirsh, N. D. M. (1930) A study of natio-racial mental differences. Genet. Psychol. Monog., Nos. 3 and 4. *716, 722*

Hoffman, G. J. (1923) An experiment in self-estimation. J. Abn. and Social Psychol., 18, 43–49. *711*

Holden, W. A., and Bosse, K. K. (1900) Order of development of color perception and color preference in the child. Arch. of Ophthal., 29, 261–278. *401*

Hollingworth, H. L. (1911) Judgments of persuasiveness. Psychol. Rev., 18, 234–256. *694, 696*

Holzinger, K. J. (1929) The relative effect of nature and nurture influences on twin differences. J. Ed. Psychol., 20, 421–428. *716*

——. (1934) Preliminary Report on Spearman-Holzinger Unitary Trait. Nos. 1 and 2. (1935) Nos. 3, 4, 5, and 6. Statist. Lab. Dept. Ed., Univ. Chicago. *339*

——. (1935 a) Statistical résumé of the Spearman two-factor theory. Univ. of Chicago Press. *339*

Holzinger, K. J., and Harman, H. H. (1937) Relationships between factors obtained from certain analyses. J. of Ed. Psychol., 28, 321–345. *324, 338, 345*

Holzinger, K. J., and Swineford, F. (1936) Preliminary report on Spearman-Holzinger unitary trait study. Nos. 7 and 8. Statist. Lab., Univ. Chicago. *339*

Hoppock, R. (1935) Job satisfaction. Harper, N. Y. Pp. xxi + 303. *452, 453, 455*

Horn, E. (1926) A basic writing vocabulary; 10,000 words most commonly used in writing. State Univ. of Iowa Monog. in Ed. Series, No. 4. *200*

Horowitz. E. L. (1936) Development of attitude toward the negro. Arch. Psychol., No. 194. Pp. 47. *486, 496, 498, 514*

Horowitz, R., and Murphy, L. B. (1938) Projective methods in the psychological study of children. J. Exp. Ed., 7, 133–140. *534*

Hotelling, H. (1933) Analysis of a complex of statistical variables into principal components. J. Ed. Psychol.. 24, 417–441. *141*

Howell. T. H. (1933) An experimental study of persistence. J. Abn. and Social Psychol., 28, 14–29. *570*

Hull. C. L. (1928) Aptitude testing. World Book Co., Yonkers, N. Y. Pp. 535. *694*

Humm, D. G., and Wadsworth, G. W. (1934) The Humm-Wadsworth temperament scale. Pers. J., 12, 314–323. *540*

Huston, P. E., Shakow, D., and Erickson, J. H. (1924) A study of hypnotically induced complexes by means of the Luria Technique. J. Gen. Psychol., 11, 65–97. *576*

Irmina, Sister M. (1928) The effects of summer vacation upon the retention of the elementary school subjects. Ed. Res. B., Cath. Univ. Am., Nos. 3 and 4. *635*

Jacobs, J. (1887) Experiments on prehension. Mind, 12, 75–79. *298*

Jacobsen, M. M. (1928) The quantitative determination of scholastic interests among college students. Unpublished Ph.D. Dissertation. Univ. of Minn. Library. Pp. 181. *479*

Jarvie, L. L., and Johns. A. A. (1938) Does the Bernreuter personality inventory contribute to counseling? Ed. Res. B., 17, 7–9. *581*

Jastak, J. (1938) School test patterns of clinic children. Delaware State Med. J., 10, 108–111. *254*

Jastrow, J. (1891–1892) Some anthropological and psychologic tests on college students — a preliminary survey. Am. J. of Psychol., 4, 420–427. *298*

Jennings, H. S. (1930) The biological basis of human nature. Norton, N. Y. Pp. 384. *713*

Jersild. A. T.. and Bienstock. S. F. (1935) Development of rhythm in young children. Child Devel. Monog., No. 22. Pp. 97. *575, 617*

Jersild. A. T.. and Meigs, M. F. (1939) Observation as a research method. Rev. Ed. Res., 9. 472–482. *35*

Johnson. K. L. (1911) M. Binet's method for the measurement of intelligence: some results. J. Exp. Ped., 1. 24–31. *268*

Johnson, P. O. (1930) The permanence of learning on elementary botany. J. Ed. Psychol., 21, 37–47. *636, 637*

JOHNSON, W. (1934) The effect of mood on personality traits as measured by Bernreuter. J. Soc. Psychol., 5, 515–522. *597*

JONES, G. A., AND LASLETT, H. R. (1935) Prediction of scholastic success in college. J. Ed. Res., 29, 266–271. *257*

JONES, H. E. (1922) Experimental studies in college teaching. Arch. of Psychol. Pp. 68. *635, 637*

——. (1928) A first study of parent-child resemblance in intelligence. 27th Yr. Book, Nat. Soc. Stud. Ed., Part 1, 61–72. *716*

——. (1937) Reaction time and motor development. Am. J. Psychol., 50, 181–194. *365*

JONES, H. E., AND BAYLEY, NANCY. (1937) Environmental correlates of mental and motor development: a cumulative study from infancy to six years. Child Devel., 8, 329–341. *719*

JONES, H. E., AND CONRAD, H. S. (1933) Growth and decline of intelligence. A study of a homogeneous group between the ages of ten and sixty. Genet. Psychol. Monog., 13, No. 3, 223–298. *651, 652*

JONES, H. E., CONRAD, H. S., AND BLANCHARD, M. B. (1932) Environmental handicap in mental test performance. Univ. of Calif. Pub. Psychol., 5, 80–83. *725*

JONES, M. C. (1926) The development of early behavior patterns in young children. Ped. Sem., 33, 537–585. *182*

JORDAN, A. M., AND VAN WAGENEN, M. J. (1933) Analytical scales of attainment in literature. Ed. Test Bur., Minneapolis. *423*

JORGENSEN, A. N., AND GREENE, H. A. (1927) Iowa high school silent reading tests. Bur. Ed. Res. and Service, Univ. of Iowa. *225*

JUNG, C. G. (1916) Psychology of the unconscious. Moffat, Yard, N. Y. Pp. 566. *524*

KANNER, L. (1935) Child psychiatry. Thomas, Springfield, Ill. Pp. 527. *151, 181*

KEFAUVER, G. N., HAND, H. C., BLOCK, V. L., AND PROCTOR, W. M. (1937) Kefauver-Hand guidance tests and inventories. World Book Co., Yonkers, N. Y. *458*

KELLETT, K. H. (1939) A Gestalt study of the function of unity in aesthetic perception. Psychol. Monog., 51, 23. *390*

KELLEY, T. L. (1916) A constructive ability test. J. Ed. Psychol., 7, 1–17. *344*

——. (1920) The Trabue scales adapted for individual testing. Teach. Col., Columbia Univ. *650*

——. (1927) Interpretation of educational measurements. World Book Co., Yonkers, N. Y. Pp. 363. *251, 257*

——. (1928) Crossroads in the mind of man. Stanford Univ. Press. Pp. 238. *141*

——. (1935) Essential traits of mental life. Harvard Stud. Ed., 26. Pp. 145. *251, 257*

KELLEY, T. L., AND KREY, A. C. (1934) Tests and measurements in the social sciences. Scribner, N. Y. Pp. 635. *201, 236, 238, 537*

KELLEY, T. L., RUCH, G. M., AND TERMAN, L. M. (1922, revised 1925, 1929, 1940) The Stanford achievement test. World Book Co., Yonkers, N. Y. *63*

KELTY, M. G., AND MOORE, N. E. (1934) Tests of concepts in the social studies. Scribners, N. Y. *238*

KENT, G. H., AND ROSANOFF, A. J. (1910) A study of association in insanity. Am. J. Insanity, 67, 37, 317. *31*

KERCHENSTEINER, D. G. (1905) Die Enwickeling der Zeichnerischen Begabung. Gerber, Munich. Pp. 508. *404*

KERR, M. (1937) Children's drawings of houses. Brit. J. Med. Psychol., 16, 206–218. *534*

King, I., and Adelstein, M. (1917) The permanence of interests and their relation to abilities. Sch. and Soc., 6, 359–360. *477*

Kingsbury, F. A. (1937) Why do people go to church? Religious Ed., 32, 50–54. *314*

Kirkpatrick, E. (1891) Number of words in an ordinary vocabulary. Sci., 18, 107–108. *298*

Kline, L. W., and Carey, G. L. (1922 and 1923) The Kline-Carey measuring scale for freehand drawing. Johns Hopkins Univ. Press, Baltimore. *406*

Klopfer, B., and Davidson, H. H. (1937) Rorschach statistics. Part I: Mentally retarded, normal, and superior adults. Rorschach Res. Exch., 2, 164–169. *528*

Knauber, A. J., and Pressey, S. L. (1927) Art ability test. Public School Pub. Co., Bloomington, Ill. *411*

Knower, F. H. (1935) Experimental studies in changes in attitudes: a study of the effect of oral argument of changes of attitude. J. Soc. Psychol., 6, 315–347. *511*

Knox, H. A. (1914) A scale based on work at Ellis Island, for estimating mental defect. J. Am. Med. Assn., 62, 741–747. *344*

Koffka, K. (1928) The growth of the mind. An introduction to child psychology. Trans. by Robert Morris Ogden. 2nd Ed., Rev. Harcourt, Brace, N. Y. Pp. 427. *181*

Kohn, H. A. (1938) Achievement and intelligence examinations correlated with each other and with teacher's rankings. J. Genet. Psychol., 52, 433–437. *257*

Kohs, S. C. (1927) Intelligence measurement. Macmillan Co., N. Y. Pp. 312. *336, 351, 619*

Koos, L. V. (1918) Determination of ultimate standards of quality in handwriting for public schools. Elem. Sch. J., 18, 423–446. *194*

Kornhauser, A. W. (1929) Results from a quantitative questionnaire on likes and dislikes with a group of college freshmen. J. Appl. Psychol., 11, 85–94. *478*

Korperth-Tippel, A. (1935) Kind und Bild Wien, Arb. Padag. Psychol., No. 10. Pp. 126. *396*

Kraepelin, Emil. (1895) Der Psychologische Versuch in der Psychiatrie. Psychologisch Arbeiten, Vol. 1. Pp. 1–91. *298*

Krasnagorsky, N. (1907) Uber die bildung der künstichen bedingungs reflexl bei Sauglingen. Russki Wratsch, 36, 1245–1246. *151*

Kroll, A. (1934) The teacher's influence upon the social attitude of boys in the twelfth grade. J. Ed. Psychol., 25, 274–280. *510*

Kuder, G. F. (1939) The stability of preference items. J. Soc. Psychol., 10, 41–50. *467*

——. (1939) Preference record. Univ. of Chicago Press. *468*

Kuhlmann, F. (1911) Binet and Simon's system for measuring the intelligence of children. J. of Psycho-Asthenics, 15, 76–92. *268*

——. (1912) A revision of the Binet-Simon system for measuring the intelligence of children. J. of Psycho-Asthenics, Monog. Suppl. *269*

——. (1921) The results of repeated mental re-examinations of 639 feebleminded over a period of ten years. J. Appl. Psychol., 5, 195–224. *294*

——. (1922) A handbook of mental tests. A further revision and extension of the Binet-Simon scale. Warwick and York, Baltimore. Pp. 208. *152, 182, 273, 291*

——. (1939) Tests of mental development. Ed. Test Bureau, Minneapolis. Pp. 314. *273, 278, 282, 289, 291*

Kuhlmann, F., and Anderson, R. (1927) Kuhlman-Anderson intelligence tests. Ed. Test Bureau, Minneapolis. *328*

Kwalwasser, J. (1926) Tests of melodic and harmonic sensitivity. Author, Syracuse Univ. P. 3. *430*

——. (1927) Tests and measurements on music. Birchard and Co., 145. *432*

——. (1930) Kwalwasser-Dykema Music Tests. Carl Fisher, N. Y. *427*

Kwalwasser, J., and Ruch, G. M. (1924) K-R Tests of musical accomplishment. Extension Div., State Univ. of Iowa. *430*

Laird, D. A. (1925) Detecting abnormal behavior. J. Abn. and Soc. Psychol., 20, 128–141. *538, 541*

Lamson, E. E. (1938) To what extent are intelligence quotients increased by children who participate in a rich, vital school curriculum? J. Ed. Psychol., 29, 67–70. *726*

Landis, C. *et al.* (1935) Empirical evaluation of three personality adjustment inventories. J. Ed. Psychol., 26, 321–330. *581*

Langfeld, H. S. (1920) The aesthetic attitude. Psychol. Monog., 51, 22. *385*

Lankes, W. Natural rate of tapping. Brit. J. Psychol., 7, 390–391. *589*

Lark-Horovitz, Betty. (1936) Interlinkage of sensory memories in relationship to training in drawing. J. of Genet. Psychol., 49, 69–89. *404*

——. (1938) On art appreciation of children: preference of picture subjects in general. J. of Ed. Res., 31, 118–137. *388, 389*

Larson, J. A., and Haney, G. W. (1932) Cardio-respiratory variations in personality studies. Am. J. Psychiat., 11, 1035–81. *578*

Larson, W. S. (1938) Practical experience with music tests. Music Educators, 24, No. 5, 31, 68–74. *436*

Lauterbach, C. E. (1928) Studies in twin resemblance. Genetics, 10, 525–568. *722*

Layman, E. McC. (1937) An item analysis of the adjustment questionnaire. Psychol. B., 34, 782. *584, 588*

Lee, J. M. (1930) Lee test of algebraic ability. Public School Pub. Co., Bloomington, Ill. *229*

Lehman, H. C., and Witty, P. A. (1927) The psychology of play activities. A. S. Barnes, N. Y. Pp. 242. *447, 658*

Levey, H. B. (1938) Poetry production as a supplemental emergency defense against anxiety. Psychoanal. Quart., 7, 232–242. *425*

Lewerenz, A. S. (1927) Test in fundamental abilities of visual arts. S. Calif. School Book Depository, Hollywood, Calif. *411*

Lewis, E. E. (1923) English composition scales for measuring business and social correspondence. World Book, Yonkers, N. Y. *198*

Lewis, M. M. (1936) Infant speech: a study of the beginnings of language. Harcourt, Brace, N. Y. Pp. 335. *179*

Liebold, R. (1936) Kind und Metronom. Z. Pedag. Psychol., 37, 317–322. *375*

Likert, R. (1932) A technique for the measurement of attitudes. Arch. Psychol., No. 140, Columbia Univ., 1–55. *518*

Likert, R., and Quasha, W. H. (1934) Revised Minnesota paper form board. Published by the authors, New York University. *357*

Lincoln, E. A. (1936) Stanford-Binet IQ changes in the Harvard growth study. J. Appl. Psychol., 20, 258–263. *663*

Line, W., and Griffin, J. D. M. (1935) Some results obtained with the Rorschach test, objectively scored. Am. J. Psychol., 92, 109–114. *590*

Linfert, H. E., and Hierholzer, H. M. (1928) A scale for measuring the mental development of infants during the first year of life. Cath. Univ. Am., Stud. in Psychol. and Psychiat., 1, No. 4. Pp. 33. *183*

Link, H. C. (1919) Employment psychology. Macmillan, N. Y. Pp. 440. *53, 359, 372*

——. (1936) A test of four personality traits of adolescents. J. of Appl. Psychol., 20, 527, 534. *515, 516*

Lorge, I. (1935) Personality traits by Fiat. II. The consistency of the Bernreuter personality inventory by the Bernreuter and by the Flanagan Keys. J. Ed. Psychol., 26, 6, 427–431. *596*

Luria, A. R. (1932) The nature of human conflicts. Liverright, N. Y. Pp. 431. *571*

Lurie, W. A. (1937) A study of Spranger's value-types by the method of factor analysis. J. Soc. Psychol., 8, 17–37. *506*

McAdory, Margaret. (1929) Art Test. The construction and validation of an art test. Contrib. to Ed., Teach. Col., Columbia Univ., No. 11. *35, 393, 395*

McCall, W. A. (1922) How to measure in education. Macmillan, New York. *683*

McCall, W. A., Herring, J. P., and Loftus, J. J. (1937) Comprehensive testing program. Laidlow Brothers, Chicago. *323*

McCarty, S. A. (1924) Children's drawings. Williams and Wilkins, Baltimore. *406*

McCaskill, C. L., and Wellman, B. L. (1938) A study of common motor achievements at the preschool ages. Child Devel., 9, 141–149. *367, 370*

McClatchy, V. R. (1928) A theoretical and statistical study of the personality trait originality as herein defined. J. Abnorm. and Social Psychol., 23, 379. *536*

McCloy, W. (1939) Creative imagination in children and adults. Psychol. Monog., 51, No. 5, 88–102. *407*

McClusky, H. Y. (1934) An experimental comparison of two methods of correcting the outcome of an examination. Sch. and Soc., 40, 566–568. *114*

McDougall, Wm. (1905, 1908) Physiological psychology. J. M. Dent, London. *22, 21*

McGraw, M. B. (1935) Growth: a study of Johnny and Jimmy. Appleton-Century Co., New York. Pp. 319. *722*

McHale, Kathryn. (1930) An information test of interests. Psychol. Clin., 29, 53–58. *470*

McNemar, Q. (1933) Twin resemblances in motor skills, and the effects of practice thereon. J. Genet. Psychol., 42, 70–99. *726*

——. (1936) Practice and 'general' motor ability. J. Gen. Psychol., 14, 464–466. *631*

——. (1939) Sampling in factor analysis. Psychol. B., 36, p. 530. *339*

Macfarlane, J. W. (1938) Methodology of data collection and organization. Monog. Soc. Res. Child Devel., 3, No. 6, 253. *169, 184*

MacQuarrie, T. W. (1925) MacQuarrie test for mechanical ability. Res. Service Co., Los Angeles, Cal. *358*

Maizlish, I. L. (1936) New possibilities in intelligence testing: interview form. J. Appl. Psychol., 20, 599–608. *320*

Maller, J. B. (1932) Character and personality tests. Teach. Col., Columbia Univ., N. Y. Pp. 53. *553, 596*

——. (1934) General and specific factors in character. J. Soc. Psychol., 5, 97–102. *592*

Manry, C. (1927) World citizenship. Univ. of Iowa Stud.: Stud. Char., I, No. 1. *515*

Manske, A. J. (1936) The reflection of teachers' attitudes in the attitudes of their pupils. Contrib. to Ed., Teach. Col., Columbia Univ., No. 702. Pp. 67. *510*

Manuel, H. T., and Hughes, L. S. (1932) The intelligence and drawing ability of young Mexican children. J. Appl. Psychol., 16, 382–387. *410*

Marple, C. H. (1933) The comparative suggestibility of three age levels to the suggestion of group vs. expert opinion. J. Social Psychol., 4, 176–186. *513*

Marsh, S. E., and Perrin, F. A. C. (1925) An experimental study of the rating scale technique. J. Abn. and Social Psychol., 19, 383–399. *709*

Maslow, A. H. (1937) Dominance feeling, behavior, and status. Psychol. Rev., 44, 404–420. *581*

Mathews, C. O. (1927) The effect of position of printed response words upon children's answers to questions in two-responses types of tests. J. Ed. Psychol., 18, 445–457. *705*

Mathews, E. (1923) A study of emotional stability in children by means of a questionnaire. J. Delinq., 8, 1–40. *538*

Meier, N. C. (1935) Diagnosis in art. 34th Yr. Book, Nat. Soc. Study Ed., 463–76. *405*

——. (1939) Studies in the psychology of art. Psychol. Monog., 51, 1–158. *413*

Meier, N. C., and Seashore, C. E. (1930) The Meier-Seashore art judgment test. Examiner's manual. Bur. Ed. Res. and Serv., Univ. of Iowa. *393, 394, 395*

Meltzer, H. (1935) Children's attitudes to parents. Am. J. Orthopsychiatry, 5, 244–265. *538*

Merrill, M. and A. (1924) On the relation of intelligence to achievement in the case of mentally retarded children. Comp. Psychol. Monog., No. 10. Pp. 100. *277, 279*

Merrill, M. A. (1938) Significance of IQ's on the revised Stanford-Binet scales. J. Ed. Psychol., 29, 641–651. *290*

Merriman, C. (1923) The intellectual resemblance of twins. Psychol. Monog., 33, No. 5, 1–58. *722*

Metcalf, J. T. (1938) Psychological studies of literary form. Psychol. B., 35, 337–357. *431*

Metcalfe, Z., and Burr, E. T. (1937) I. E. R. assembly test revised norms based on short form. J. Appl. Psychol., 21, 372–378. *361*

Meumann, E. (1905) Intelligenzprufung un der Kinder der Volpechule. Experimentelle Pedagogik, 1, 35–100. *269*

Meyer, G. (1936) The effect of recall and recognition of the examination set in classroom situation. J. Ed. Psychol., 27, 81–99. *113*

Miles, W. R. (1922) Notes on electric counters. J. Exp. Psychol., 51, 76–77. *373*

——. (1931) Measurement of certain human abilities throughout the life span. Proc. Nat. Acad. Sci., 17, 627–633. *365, 647, 649*

——. (1933) Age and human ability. Psychol. Rev., 40, 99–123. *649*

Miles, W. R., and Miles, C. S. (1932) The correlation of intelligence scores and chronological age from early to late maturity. Am. J. Psychol., 44, 44–78. *653, 654*

Mirenova, A. N. (1935) Psychomotor education and the general development of pre-school children: experiments with twin controls. J. Genet. Psychol., 46, 433–454. *722*

Monroe, Marian. (1935) Reading aptitude tests. Houghton Mifflin, Boston. *172*

Moore, H. K. (1937) Confession tests. J. Juv. Res., 21, 76–83. *597*

Moore, H. T. (1914) The genetic aspect of consonance and dissonance. Psychol. Monog., No. 73, 1–68. *428*

——. (1921) The comparative influence of majority and expert opinion. Am. J. Psychol., 32, 16–20. *513*

Moos, J. C. (1930) The yardstick applied to musical talent. Mus. Q., 16, 238, 261. *436*

Morgan, T. H. (1928) The mechanism and laws of heredity. The International University Series, Clark Univ. Press, Worcester, Mass. *638*

Morgan, C. D., and Murray, H. H. (1935) A method for investigating fantasies: the thematic apperception test. Arch. Neur. Psychiat., 34, 289–306. *533*

Morris, O. M. (1939) A critical analysis of certain performance tests. J. Genet. Psychol., 54, 85–106. *382*

Mosher, R. M. (1928) A further note on the reliability of reading tests. J. Ed. Psychol., 19, 272–274. *617, 618*

Mosier, C. I. (1937) A factor analysis of certain neurotic tendencies. Psychometrika, 2, 263–287. *538, 539, 585, 586*

Moss, F. A. (1931) Development of the Wabblemeter for measuring bodily steadiness as an indication of riding comfort. J. Soc. of Automotive Engineers, 28, 5, 580–581. *374*

Muller, H. F. (1925) Mental traits and heredity. J. of Hered., 16, 433–448. *722*

Murphy, G., and Likert, R. (1938) Public opinion and the individual. A psychological study of student attitudes on public questions, with a retest five years later. Harper, New York. Pp. 316. *486, 493, 495, 512, 513, 514, 515, 517*

Murphy, G., Murphy, L. B., and Newcomb, T. M. (1937) Experimental social psychology. Harper, New York. Pp. 1119. *511, 520, 596*

Mursell, J. L. (1937) What about music tests? Music Ed. J., 24. *437*

Musico, B. (1922) Motor capacity with special reference to vocational guidance. Brit. J. Psychol., 13, 157–184. *372, 379*

Myers, C. S. (1927) The effects of music, a series of essays edited by M. Shoen. Harcourt, Brace, Section 1. Pp. 273. *429*

Nemzek, C. L. (1933) The constancy of the I.Q. J. Appl. Psychol., 17, 475–77. *294*

Netzer, R. L. (1938) Stimuli for oral language. Elem. Engl. Rev., 15, 91–94. *198*

Newcomb, T. M. (1929) The consistency of certain extrovert-introvert behavior patterns in 51 problem boys. Contrib. to Ed., Teach. Col., Columbia Univ., No. 382. *562, 563*

——. (1931) An experiment designed to test the validity of a rating technique. J. Educ. Psychol., 22, 279–289. *707*

Newman, H. H. (1933) Mental and physical traits of identical twins reared apart. J. Heredity, 20, 49–64, 97–104, 153–166. *722*

Newman, H. H., Freeman, F. N., and Holzinger, K. J. (1937) Twins: a study of heredity and environment. Univ. Chicago Press. Pp. 369. *713*

Newstetter, W. I. (1937) An experiment in the defining and measuring of group adjustment. Amer. Sociol. Rev., 2, 230–236. *582*

Noonan, M. E. (1926) Influence of summer vacation on the abilities of fifth and sixth grade children. Contrib. to Ed., Teach. Col., Columbia Univ., No. 204. *636*

Nystrom, G. (1930) Minneapolis self-corrective handwriting charts. St. Paul Book and Stationery Co., St. Paul. *194*

O'Brien, J. A. (1921) Silent reading with special reference to methods for developing speed. Macmillan, N. Y. Pp. 289. *611*

O'Connor, J. (1928) Born that way. Williams and Wilkins, Baltimore, Md. Pp. 323. *359, 373*

——. (1934) Psychometrics. Harvard Univ. Press. Pp. 64. *536*

Odom, C. L. (1929) A study of the mental growth curve with special reference to the results of group intelligence tests. J. Ed. Psychol., 20, 401–416. *683, 686, 688*

Oehrn, A. (1895) Experimentelle Studien zur Individualpsychologie. Psychologischen Arbeiten, 1, 92–152. *298*

Olson, D. M., and Jones, Vernon. (1931) An objective measure of emotionally toned attitudes. Ped. Sem. and J. Genet. Psychol., 39, 174–196. *576*

Olson, W. C. (1936) The waiver of signature in personal reports. J. Appl. Psychol., 20, 442–450. *597*

——. (1937) The concept of the organism as a whole: interrelationships in physical, mental, social, and emotional development. Bur. Educ. Ref. and Res., Univ. of Mich. Pp. 8. *665*

Olson, W. C., and Cunningham, E. M. (1934) Time sampling techniques. Child Devel., 5, 41–58. *569*

Orleans, J. B., and Orleans, J. S. (1934) A study of prognosis in high school algebra. Math. Teach., No. 22, 23–30. *257*

O'Rourke, L. J. (1932) O'Rourke general classification test. Psychol. Inst., Washington, D. C. *321*

——. (1934) Rebuilding the English usage curriculum. Psychol. Inst., Washington, D. C. Pp. 98. *211*

——. (1935) O'Rourke clerical aptitudes test, junior grade. Psychol. Inst., Washington, D. C. *33, 322*

——. (1936) O'Rourke mechanical aptitude test, junior grade, form c. Psychol. Inst., Washington, D. C. *356*

Ortmann, O. (1929) The physiological mechanics of piano technique. Dutton, N. Y. *427*

Otis, A. S. (1918) Otis group intelligence scale. World Book Company, Yonkers, N. Y. *320*

——. (1922) Otis Self-Administering Tests of Mental Ability. World Book Co., Yonkers, N. Y. *321*

Parker, D. H. (1926) The analysis of art. Yale Univ. Press. Pp. 190. *385, 386*

Paterson, D. G., and Darley, J. G. (1936) Men, women, and jobs. U. of Minn. Press. Pp. 145. *93*

Paterson, D. G., Elliott, R. M., Anderson, L. D., Toops, H. A., and Heidbreder, Edna. (1930) Minnesota mechanical ability tests. Univ. of Minn. Press. Pp. 586. *28, 354, 355, 359, 360, 371, 373, 374, 376, 380, 619*

Patrick, C. (1937) Creative thought in artists. J. Psychol., 4, 35–73. *404*

Pearl, R., and Reed, L. J. (1920) On the rate of growth of the population of the United States since 1790, and its mathematical representation. Proc. Nat. Acad. Sci., Vol. y, 275–288. *670*

Pearson, K. (1904) On the laws of inheritance in man. Biometrika, 3, 131–190. *65*

——. (1914) Tables for statisticians and biometricians. Cambridge Univ. Press, England. Pp. 143. *60*

——. (1915) Some recent misinterpretations of the problem of nature and nurture. Eugenics Lab. Lec. Series. III, Part 2. *716*

Perrin, F. A. C. (1921) An experimental study of motor ability. J. Exp. Psychol., 4, 24–56. *379*

Personnel Manual. (1919) Committee on Classification of Personnel. Adjutant General's Dept., Washington, D. C. C. C. P. 400. Pp. 342. *702*

Peterson, Joseph. (1917) Experiments in ball tossing: the significance of learning curves. J. of Exp. Psychol., 2, 178–224. *623*

——. (1925) Early conceptions and tests of intelligence. World Book Co., Yonkers, N. Y. Pp. XIV + 298. *267*

Peterson, J., and Barlow, M. C. (1928) The effects of practice on individual differences. 27th Yr. Book, Nat. Soc. Stud. Educ., Part II, 211–230. *623*

Peterson, T. J. (1938) A preliminary study of the effects of previous nursery school attendance upon five-year-old children entering kindergarten. Univ. Ia. Stud. Child Welf., 14, 197–218. *726*

Petran, L. A. (1937) An analysis of the Ortmann tests in tonal discrimination and memory. B. Amer. Musicol. Soc., No. 2, 25. *434, 436*

Piaget, Jean. (1926) The language and thought of the child. Trans. by Marjorie Warden. Harcourt, Brace, N. Y. Pp. xxiii + 246. *180*

——. (1929) The child's conception of the world. Harcourt, Brace, N. Y. Pp. ix + 397. *181*

——. (1930) The child's conception of physical causality. Trans. by Marjorie Gabain. Harcourt, Brace, N. Y. Pp. 309. *181*

Pintner, R. (1918) The mental indices of siblings. Psych. Rev., 25, 252–255. *716*

——. (1918 a) Aesthetic appreciation of pictures by children. Ped. Sem., 25, 216–220. *396*

——. (1919) A non-language group intelligence test. J. of Appl. Psychol., 3, 199–214. *352*

——. (1931) Intelligence testing, methods and results. Henry Holt, N. Y. Pp. 555. *291*

Pintner, R., and Hildreth, G. (1937) Pintner-Paterson performance tests, short scale; ages 4 to 15. Bur. of Pub. Teach. Col., Columbia Univ. *350*

Pintner, R., and Paterson, D. G. (1915) The Binet scale and the deaf child. J. of Ed. Psychol., 6, 591–600. *344*

——. (1917) A scale of performance tests. Appleton-Century, N. Y. *336, 344*

Pintner, R., and Renshaw, S. (1920) A standardization and weighting of two hundred analogies. J. of Appl. Psychol., 4, 263–273. *650*

Poffenberger, A. T., and Barrows, B. E. (1924) The feeling value of lines. J. of Appl. Psychol., 1, 187. *397, 398*

Porteus, S. D. (1924) Guide to Porteus maze test. Vineland Train. School B., No. 25 (Original Scale Published in 1915). *350*

Powers, S. R. (1921) Achievement in high-school chemistry: an examination of subject matter. Teach. Col. Rec., 15, 203–211. *636*

Pratt, C. C. (1931) The meaning of music. McGraw-Hill, N. Y. Pp. 253. *436*

Pressey, L. C. (1924) An investigation of the technical vocabularies of the school subjects. Ed. Res. B. Ohio State Univ., 3, 182–185. *202*

——. (1934) Tests and measurements in the social sciences, Chapter III. Scribners, N. Y. Pp. 155. *201, 204, 205, 237*

Pressey, L. C. and S. L. (1922) Intermediate classification test. Public School Pub. Co., Bloomington. Ill. *321*

Pressey, S. L. (1933) Psychology and the new education. Harper, N. Y. Pp. 594. *249, 260*

Pressey, S. L. and L. M. (1918) Pressey's group point scale for measuring general intelligence. J. of Appl. Psychol., 2. 250–259. *320*

Proctor, W. M. (1923) Psychological tests and guidance of high school pupils. J. Ed. Res. Monog., No. 1. Pp. 125. *448*

Purdom, L. T. (1932) Purdom diagnostic blank. Author, Ann Arbor. Mich. *545*

Pyle, W. H. (1913) The examination of school children. Macmillan Co., N. Y. Pp. 70. *300*

Rand, G. (1925) A discussion of the quotient method of specifying test results. J. of Ed. Psychol., 16. 599–618. *251*

Randall, H. M., Williams, N. H., and Colby, W. F. (1929) General college physics. Harper. N. Y. Pp. 539. *668*

Raths. L. E. (1938) Appraising certain aspects of student achievement. Nat. Soc. Stud. Ed. Yr. Book. Chap. 3, 89–117. *92*

Ream, M. J. (1924) Ability to sell: its relation to certain aspects of personality and experience. Williams and Wilkins, Baltimore. Pp. 64. *459, 460*

Reckless, W. C., and Bringen, H. L. (1933) Racial attitudes and information about the negro. J. Negro Ed., 2, 128–138. *315*

Reed, H. B. (1931) The influence of training on changes in variability in achievement. Psychol. Monog., 41. Pp. 59. *623, 628*

Remmers, H. H. (1929) The measurement of interest differences between students of engineering and agriculture. J. of Appl. Psychol., 13, 105–119. *480*

Remmers, H. H., and Silance, E. F. (1934) Generalized attitude scales. J. of Social Psychol., 5, 298–312. *697*

Richardson, M. W., and Stalnaker, J. M. (1935) Comments on achievement examinations. J. Ed. Res., 28, 25–32. *257*

Richardson, C. A., and Stokes, C. W. (1935) Growth and variability of intelligence. Brit. J. of Psychol. Monog. Supplements, No. 18. Cambridge U. Press, England. Pp. 83. *677*

Rigg, M. (1937) Rigg poetry test, forms C and D. Sch. of Ed., Stillwater, Okla. *422*

——. (1937) An experiment to determine how accurately college students can interpret the intended meanings of musical compositions. J. Exp. Psychol., 21, 223–229. *429*

Robertson, T. B. (1913) On the nature of the autocatalyst of growth. Arch. f. Entwicklungsmech. der Organismen, 37. Pp. 497. *670*

Roblee, L., and Washburn, M. F. (1912) The affective value of articulate sounds. Am. J. Psychol., 23, 579–583. *419*

Rodger, A. G. (1936) The application of six group intelligence tests to the same children, and the effects of practice. Brit. J. Ed. Psychol., 6, 291–305. *615*

Rorschach, H. (1921) Psychodiagnostik: Methodik und Ergebnisse eines wahrehmungs-diagnostischen Experiments. Pp. 174. (1932) Second Edition, Huber, Bern, Switzerland. Pp. 227. *32, 524*

Rosanoff, A. J. (1920) Manual of psychiatry. Wiley and Sons, N. Y. Pp. 684. *31, 534*

Rousseau, J. J. (1762) Emile, or education. Dutton, N. Y. Pp. 444. *151*

Ruch, G. M., Cleeton, G. U., and Stoddard, G. D. (1925) Iowa High School Content Examination. Copyright by G. M. Ruch. *235*

Ruch, G. M., Knight, F. B., Greene, H. A., and Studebaker, J. W. (1925) The compass diagnostic tests in arithmetic. Scott, Foresman, Chicago. *228*

Ruch, G. M., and Stoddard, G. D. (1925) Comparative reliabilities of five types of objective examinations. J. Ed. Psychol., 16, 89–103. *125, 126*

——. (1927) Tests and measurements in high school instruction. World Book Co., Yonkers, N. Y. Pp. 381. *115, 127, 128*

Ruckmick, C. A. (1938) A critique of the galvanic technique. Psychol. Rev., 45, 154–162. *578*

Rundquist, E. A., and Sletto, R. F. (1936) Minnesota scale for the survey of opinion. Univ. of Minn. Press. *520, 541, 542, 543*

——. (1936) Personality in the depression. Univ. of Minn. Press. Pp. 398. *596, 597*

Salisbury, F. S., and Smith, H. B. (1929) Prognosis of sight singing ability. J. of Appl. Psychol., 13, 425–439. *433*

Sanford, E. C. (1908) A course in experimental psychology. Heath, Boston. Pp. 449. *696*

Sangren, P. V., and Woody, C. (1927) Sangren-Woody reading test. World Book Co., Yonkers, N. Y. *220, 221*

Santayana, G. (1896) The sense of beauty. Charles Scribner's Sons, N. Y. Pp. 275. *385*

Saunders, A. W. (1936) The stability of artistic aptitude at the childhood level. Psychol. Monog., 48, 126–154. *406*

Scates, D. E., and Fauntleroy, V. (1938) Effect of weights on certain index numbers. J. Exp. Ed., 6, 282–306. *64*

Schiller, B. (1934) Verbal, numerical, and spatial abilities of young children. Arch. of Psychol., No. 161, 1–69. *253*

Schneck, M. M. R. (1929) The measurement of verbal and numerical abilities. Arch. Psychol., No. 107. Pp. 49. *253*

Schoen, M. (1928) The esthetic attitude in music. Iowa Studies in Psychol., 12, 162–183. *429, 430*

Schorling, R., and Sanford, V. (1926) Achievement test in plane geometry. Bur. of Publications, Teach. Col., Columbia Univ., N. Y. *231, 232*

Schott, E. L. (1937) I.Q. changes in foster home children. J. of Appl. Psychol., 21, 107–120. *722*

Schramm, W. L. (1935) Approaches to a science of English verse. U. of Iowa. Stud. Service, Aims, and Progress Res., No. 46. Pp. 82. *432*

Schwartz, Louis A. (1931) Schwartz social situation picture test. C. H. Stoelting Co., Chicago. *533*

Schwesinger, G. C. (1933) Heredity and environment. Macmillan, N. Y. Pp. 484. *713*

Scripture, E. W. (1902) Researches in experimental phonetics. Yale Psychol. Stud., 10, 49–80. *431*

Seashore, C. E. (1919) Psychology of musical talent. Silver Burdett, N. Y. Pp. 288. *17, 24, 375, 426, 427, 435, 618*

——. (1927) Phonophotography in the measurement of expression of emotion in music and speech. Sci. Mo., 24, 463–471. *431*

——. (1938) Psychology of music. McGraw-Hill, N. Y. Pp. 408. *433*

Seashore, C. E., Lewis, D., and Saetveit, J. G. (1939) Measures of musical talents. R. C. A. Manufacturing Co., Camden, N. J. *426*

Seashore, R. H. (1926) Studies in motor rhythm. Iowa Stud. Psychol., 9, 142–199. *427*

——. (1928) Stamford motor skills unit. Psychol. Monog., 39, 51–66. *375, 377*

——. (1930) Individual differences in motor skills. J. of Gen. Psychol., 3, 38–66. *380*

Seashore, R. H., and Adams, R. D. (1933) Measurements of steadiness, a new apparatus, and results on marksmanship. Sci. 78, 285–287. *374*

Seashore, R. H., and Hevner, K. (1933) A time-saving device for the construction of attitude scales. J. Soc. Psychol., 4, 366–372. *699*

Seguin, E. (1846) (1906) Traitement moral, hygiene et education des idiots. Alcan. Paris. Pp. 20. *344*

Sharp, D. L. (1938) Group and individual profiles in the association-motor test. Univ. Ia. Stud. Child Welf., 15, No. 1. 97, 171. *577*

Sheldon, S. S. (1940) The varieties of human physique. Harper, N. Y. Pp. 347. *522*

Shen, E. (1925) The reliability coefficient of personal ratings. J. Ed. Psychol., 17, 232–236. *711*

Shepard, J. F. (1913) Keyboard puzzle box and apparatus for experimental phonetics. Psychol. B., 10, 66–67. *431*

Shirley, Mary M. (1933) The first two years. A study of 25 babies. Vol. II. Intellectual Development. Univ. of Minn. Press. Pp. xi+228. *177*

Shull, A. F. (1938) Heredity. 3rd Ed. McGraw-Hill, N. Y. Pp. 442. *713*

Shuttleworth, F. K. (1929) Measurement of the character factors other than intelligence involved in scholastic success of freshman men. Unpublished dissertation. State Univ. of Iowa. *478*

Sicha, Karel and Mary. (1937) A step toward the standardization of the scoring of the Rorschach Test. Summary in the Rorschach Research Exchange. 1, 95–101. *528*

Skeels, H. M. (1938) Mental development of children in foster homes. J. of Consulting Psychol., 2, 33–43. *721*

Skodak, M. P. (1938) The mental development of adopted children whose true mothers are feebleminded. Child Devel., 9, 303–308. *721*

Smith, R. B. (1932) The development of an inventory for the measurement of inferiority feelings at the high school level. Arch. Psychol., No. 144. Pp. 108. *596*

Snedden, D. S. (1927) Study in disguised intelligence tests. Contrib. to Ed., Teach. Col., Columbia Univ., No. 291. Pp. 48. *320*

Spearman, C. (1904) The proof and measurement of association between two things. Amer. J. Psychol., 15, 72–101. *138*

——. (1910) Correlation calculated from faulty data. Brit. J. of Psychol., 3, 281. *101*

——. (1927) The abilities of man. Macmillan, N. Y. Pp. 415. *140*

——. (1938) Psychology through the ages. 2 Vol. Macmillan, N. Y. Pp. 454, 355. *138, 590, 592*

Spencer, D. (1938) The frankness of subjects on personality measures. J. Ed. Psychol., 29, 26–35. *597*

Spranger, E. (1928) Types of men. The psychology and ethics of personality. Niemeyer, Halle. Pp. 402. *501*

Stagner, R., and Drought, N. (1935) Measuring children's attitudes toward their parents. J. Ed. Psychol., 26, 3, 164–76. *697*

Stalnaker, John M. (1935) Cream of the crop in English; an analysis of the performance of scholarship students on an English examination. Engl. J. (Col. Ed.), 466–476. *416, 417*

——. (1936) A study of optional questions on examinations. Sch. and Soc., 44, 829–832. *198*

——. (1937) Essay examinations reliably read. Sch. and Soc., 46, 671–672. *199*

Stanton, H. M. (1929) Prognosis of musical achievement. Stud. Psychol. Vol. 1. No. 4. Eastman School of Music, Rochester, N. Y. Pp. 89. *437*

Staples, R. (1932) The responsiveness of infants to color. J. Exp. Psychol., 15, 119–141. *401*

Stefanescu-Goanga, F. (1911) Experimentelle Untersuchungen zur Gefuehlsbetonung der Farben. Psychol. Studien, 7, 284–335. *399*

Steggerda, M. (1936) McAdory art test applied to Navaho Indian children. J. Comp. Psychol., 22, 283–285. *395*

Stenquist, J. L. (1923) Measurements of mechanical ability. Contrib. to Ed., Teach. Col., Columbia Univ., No. 130. *360, 619*

Stern, W. (1900) Uber Psychologie der individuellen Differenzen. J. A. Barth, Leipzig. 146. *269*

——. (1914) The psychological methods of testing intelligence (trans. by Whipple). Ed. Psychol. Monog., No. 13. Pp. 160. *269*

——. (1914 a, revised 1930) Psychology of early childhood. Henry Holt, N. Y. Pp. 623. *181*

Stewart, F. J., and Brainard, P. P. (1932) Specific interest inventory. Psychol. Corp., N. Y. *473*

Stewart, M. (1934) A scale for measuring the quality of conventional news stories in high school journalism. Engl. J., 23, 209–215. *198*

Stinchfield, S. M. (1923) The formulation and standardization of a series of graded speech tests. Psychol. Monog., 33, 149, 54. *193*

Stolz, A., and Manuel, H. T. (1931)

Art ability of Mexican children. Sch. and Soc., 31, 379–380. *395*

Stouffer, S. A. (1930) An experimental comparison of statistical and case history methods of attitude research. Ph.D. thesis, Univ. of Chicago Library. *516*

Strong, E. K. (1911) The relative merits of advertisements. Arch. Psychol., No. 17. Pp. 81. *694*

Strong, E. K., Jr. (1931) Vocational interest blank for men. Stanford Univ. Press. *9, 48*

——. (1931) Change of interests with age. Stanford Univ. Press. Pp. 235. *475*

——. (1934) Classification of occupations by interest. Personnel J., 12, 301–313. *471, 477*

——. (1935) Prediction value of vocational interest test. J. Ed. Psychol., 26, 331–349. *477*

——. (1938) Vocational interest blank for men. Stanford Univ. Press. *458, 460, 463, 464, 465*

Stump, N. F. (1935) The experimental development of an auditory group test of intelligence. Author, Keuka, N. Y. Pp. 5. *320*

Stutsman, Rachel, Ph.D. (1931) Mental measurement of preschool children. World Book Co., Yonkers, N. Y. Pp. 368. *152, 162, 163, 657*

Sullivan, E. T., Clark, W. W., and Tiegs, E. W. (1937) California test of mental maturity. Southern California Book Depository. Pp. 16. *323*

Swift, E. G. (1910) Relearning a skillful act: an experimental study in memo-muscular memory. Psychol. B., 7, 17–18. *636*

Swift, E. J. (1903) Standards of efficiency in school and in life. Ped. Seminary, 10, 3–22. *298*

Symonds, P. M. (1924) On the loss of reliability in ratings due to coarseness of the scale. J. Exp. Psychol., 7, 456–461. *704*

——. (1931) Diagnosing personality and conduct. Appleton-Century, N. Y. Pp. 602. *704, 708*

Terman, L. M. (1906) Genius and stupidity: a study of some of the intellectual processes of seven 'bright' and seven 'stupid' boys. Ped. Seminary, 13, 307–373. *267*

——. (1916) The measurement of intelligence: an explanation of a complete guide for the use of the Stanford revision and extension of the Binet-Simon intelligence scale. Houghton Mifflin, Boston. Pp. 362. *152, 203, 269, 285, 291, 310*

——. (1925) Genetic studies of genius. Vol. I. Mental and physical traits of a thousand gifted children. Stanford Univ. Press. Pp. 648. *478*

——. (1926) Terman group test of mental ability. World Book Co., Yonkers, N. Y. *320*

——. (1930) Chapter on talent and genius in children in the new generation. Macaulay, N. Y. Pp. 717. *714*

Terman, L. M., and Childs, H. G. (1912) Tentative revision and extension of the Binet-Simon measuring scale of intelligence. J. Ed. Psychol., 3, 61–74, 133–143, 198–208, 277–289. *268*

Terman, L. M., *et al.* (1917) Stanford revision and extension of the Binet-Simon scale. Ed. Psych. Monog., No. 18. Pp. 179. *103, 279, 284*

Terman, L. M., and Merrill, M. A. (1937) Measuring intelligence. Houghton Mifflin Co., Boston. Pp. 460. *24, 76, 271, 272, 278, 286, 289, 293, 294, 337, 614, 687*

Terman, L. M., and Miles, C. C. (1936) Sex and personality. McGraw-Hill, N. Y. Pp. 600. *659*

Thom, D. A. (1922) Habit clinics for children of preschool age. Mental Hygiene, 6, 463–470. *151*

Thompson, J. M. (1937) Thompson business practice test. World Book Co., Yonkers, N. Y. *245*

Thomson, W. A. (1938) An evaluation of the P. Q. test. Character and Personality, 6, 274–291. *548*

Thorndike, E. L. (1908) The effect of practice in the case of a purely intellectual function. Am. J. of Psychol., 19, 374–381. *623*

——. (1910) Handwriting. Teach. Col. Rec., 11, No. 2. *194, 696*

——. (1912) The permanence of interests and their relation to abilities. Pop. Sci. Mo., 81, 449–456. *473, 477*

——. (1916) Tests of esthetic appreciation. J. Ed. Psychol., 7, 509–522. *406*

——. (1917) Early interests: their permanence and relation to abilities. Sch. and Soc., 27, 550–565. *473, 477*

——. (1920) Intelligence examinations for college entrance. J. Ed. Res., 1, 329–337. *321*

——. (1921) The correlation between interests and abilities in college courses. Psychol. Rev., 28, 374–376. *477*

——. (1922) Practice effects on intelligence tests. J. Exp. Psychol., 5, 101–107. *614*

——. (1926) The measurement of intelligence. Bur. of Pub., Teach. Col., Columbia Univ. Pp. 616. *105, 203, 253, 328, 677, 683, 689*

——. (1927) The I.E.R. intelligence scale CAVD. Teach. Col. Bur. of Pub., Columbia Univ. *200, 716*

——. (1928) Adult learning. Macmillan, N. Y. Pp. 335. *673*

——. (1935) Adult interests. Macmillan, N. Y. Pp. 265. *474*

——. (1935 a) Thorndike-Century Junior Dictionary. Scott, Foresman, Chicago. Pp. 970. *200*

——. (1935 b) Workers' satisfactors. Likes and dislikes of young people for their jobs. Occupations, 13, 704–706. *445*

Thorndike, E. L., *et al.* (1928) The resemblance of siblings in intelligence. 27th Yr. Book, Nat. Soc. Stud. Ed., Part 1, 41–53. *716*

——. (1934) Predication of vocational success. Commonwealth Fund, N. Y. Pp. 284. *261, 452, 481*

Thorndike, E. L., Cobb, M. V., Orleans, J. S., Symonds, P. M., Wald, E., and Woodyard, E. (1923) The psychology of algebra. Macmillan, N. Y. Pp. 483. *229*

Thorndike, E. L., Woodyard, E., and Lorge, I. (1935) Four new forms of the I.E.R. intelligence scale for use on the college or higher levels. Sch. and Soc., 42, 271–272. *328*

Thorndike, R. L. (1936) Factor analysis of social and abstract intelligence. J. Ed. Psychol., 27, 231–233. *588*

Thurstone, L. L. (1919) The learning curve equation. Psychol. Monog., 20, No. 114. *670*

——. (1919) (1940) American Council on Education. Psychological examination for college freshmen. Published annually. Am. Council on Ed., Washington, D. C. *321*

——. (1921) Business information Form A. Bureau of Personnel Research, Carnegie Inst. of Tech., Pittsburgh. *245*

——. (1922) Examination in clerical work: Form A, Thurstone employment tests. World Book Co., Yonkers, N. Y. *242, 244*

——. (1925) A method of scaling psychological and educational tests. J. Ed. Psychol., 16, 433–451. *164, 369, 676, 678, 679, 680, 681, 685, 686, 688*

——. (1927) The method of paired comparisons for social values. J. Abn. and Soc. Psychol., 21, 384–400. *692*

——. (1928) The absolute zero in intelligence measurement. Psychol. Rev., 35, 175–197. *630, 643, 686*

——. (1928 a) The measurement of opinion. J. Abn. and Soc. Psychol., 22, 415–430. *693, 696*

——. (1930) The learning function. J. Gen. Psychol., 3, 469–493. *608, 609*

——. (1930 a) A neurotic inventory. J. Social Psychol., 1, 3–31. *538, 541*

——. (1931) The measurement of

change in social attitudes. J. Soc. Psychol., 2, 230–235; Influence of motion pictures on children's attitudes. J. Soc. Psychol., 2, 291–305. *512*

——. (1931 a) The reliability and validity of tests. Edwards Brothers, Ann Arbor, Mich. *124*

——. (1931 b) A multiple factor study of vocational interests. Pers. J., 10, 198–205. *470*

——. (1935) Vocational interest schedule. Psychol. Corp., N. Y. *338, 467*

——. (1935) Vectors of the mind. Univ. Chicago Press. Pp. 206. *141*

——. (1938) Primary mental abilities. Psychometrika Monog. Pp. 121. *216, 329, 339, 414*

Thurstone, L. L., and Ackerson, L. (1929) The mental growth curve for the Binet tests. J. Ed. Psychol., 20, 569–583. *683, 688, 689*

Thurstone, L. L., and Chave, E. J. (1929) The measurement of attitude, a psychophysical method. Univ. of Chicago Press. Pp. 96. *36, 486, 491, 517*

Tiebout, C. (1933) The psychophysical functions differentiating artistically superior from artistically inferior children. Psychol. Monog., 45, 108–133. *412*

Tiebout, C., and Meier, N. C. (1936) Artistic ability and general intelligence. Psychol. Monog., 48, 95–125. *395, 406, 407, 409, 410*

Tiegs, E. W., and Clark, W. W. (1934) Progressive achievement tests — advanced Battery Form A. Southern California School Book Depository, Los Angeles. *206, 212, 214, 215*

Tilton, J. W. (1936) The effect of forgetting upon individual differences. Psychol. Monog., 47, 173–185. *636*

Titchener, E. B. (1902) Ein Versuch die Methode der paarweisen Vergleichung auf die verschiedenen Gefühlsrichtungen anzuwenden. Phil. Stud., 20, 382–406. *692*

Tolstoy, L. N. (1899) What is art? Crowell, New York. *385*

Tomlinson, R. R. (1934) Picture making in children. Studio Pub. Co., N. Y. Pp. 120. *405*

Toops, H. A. (1921) Trade tests in education. Contrib. to Ed., Teach. Col., Columbia Univ., No. 115. *129*

——. (1923) Tests for vocational guidance of children thirteen to sixteen. Contrib. to Ed., Teach. Col., Columbia Univ., No. 136. *361*

——. (1937) Ohio State college entrance examination. Ohio College Assn., B., No. 90. Ohio State Univ., Columbus, O. *132, 321*

Torgenson, T. L. (1933) Public school achievement tests: geography. Public Sch. Pub. Co., Bloomington, Ill. *239*

Trabue, Marion Rex. (1916) Completion-test language scales. Teach. Col., Columbia Univ. Pp. 118. *650*

——. (1923) Scales for measuring judgment of orchestral music. J. Ed. Psychol., 14, 545–561. *430*

Travis, Lee. (1935) Educational Diagnosis. 34th Yr. Book, Nat. Soc. St. Ed. Public Sch. Pub. Co., Bloomington, Ill. Pp. 563. *192*

Traxler, A. E., and Anderson, H. A. (1935) The reliability of an essay test in English. Sch. Rev., 43, 534–539. *199*

Treves, Z., and Soffiotti, F. U. (1909, 1911) La 'Scala Metrica dell' Intelligenza di Binet e Simon. Stabilimento Tipolito grafico G. Cavelli, Milano. *268*

Trow, W. C. (1923) The psychology of confidence. Arch. of Psychol., No. 67. *130*

Tryon, R. C. (1929) Genetics of learning ability in rats. Univ. of Calif. Pub. Psychol., 4, 71–89. *717*

Uhrbrock, R. S. (1934) Attitudes of 4,430 employees. J. Social Psychol., 5, 365–377. *451*

Uhrbrock, R. S., and Richardson, M. W. (1933) Item analysis. Pers. J., 12, 141–154. *104*

United States General's Office. (1919) The personnel system in the United States army. Govt. Print. Office, Washington. *702*

Upshall, C. C. (1920) Day schools vs. institutions for the deaf. Contrib. to Ed., Teach Col., Columbia Univ. Pp. 389. *725*

Van Allesh, G. J. (1925) De Aesthetische Erscheinungsweise der Farbe. Psychol. Forshung VI, 1–91, 215–281. *400*

Van Alstyne, Dorothy. (1932) Play behavior and choice of play materials of pre-school children. Chicago, Ill.: Univ. of Chicago Press. Pp. 104. *176, 177*

Van Alstyne, D., and Osborne, E. (1937) Rhythmic responses of negro and white children two to six: with a special focus on regulated and free rhythm situations. Monog. Soc. Res. Child Devel. Pp. iv + 64. *375*

Vance, T. F., and McCall, L. T. (1934) Children's preferences among play materials as determined by the method of paired comparison of pictures. Child Devel., 5, 267–277. *533*

Van Essen, J. (1935) Bewegungsbereitschaft und netzhautbeleuchtung. Z. Sinnesphysiol. 66, 146–164. *365*

Van Wagenen, M. J. (1923) English composition scales. World Book Co., Yonkers, N. Y. *195, 197*

——. (1933) Unit scales of attainment. *Reading*, Grades VII–VIII, Form A. Ed. Test Bureau, Minneapolis. *219, 220*

——. (1934) Unit scales of aptitude. Ed. Test Bureau, Minneapolis, 3–31. *683*

Vaughn, J., and Krug, O. (1938) The analytic character of the Rorschach ink-blot tests. Am. J. of Orthopsychiatry, 8, 220–229. *529*

Vernon, M. D. (1938) The drives which determine the change of a career. Brit. J. of Ed. Psychol., 8, 1–15. *584*

Vernon, P. E. (1933) The Rorschach ink-blot test III. Brit. J. Med. Psychol., 13, 89–114, 179–205, 271–291. *527*

——. (1938) The assessment of psychological qualities by verbal methods. Medical Research Council, Industrial Health Research Board, Report No. 83. H. M. Stationary Office, London. Pp. 124. *448, 575, 588*

Vincent, L. M. (1924) A study of intelligence test elements. Contrib. to Ed., Teach. Col., Columbia Univ., No. 152. *104*

Voss, M. D. (1936) A study of conditions affecting the functioning of the art appreciation process at the child-level. Psychol. Monog., 48, 1–39. *396*

Wallar, G. A., and Pressey, S. L. (1937) Occupational orientation inventory. Authors, Ohio State University. *458, 460, 461*

Wallis, Nell. (1930) A study of tests designed to measure art capacities. Master's Thesis, Florida State College for Women, Tallahassee, Florida, 76. *395*

Walton, W. E. (1933) The sensitivity of children and adults to color harmony. Psychol. Monog., 46, 51–61. *401*

——. (1936) Exploratory experiments in color discrimination. J. Genet. Psychol., 48, 221–226. *399*

Washburne, C. (1939) The work of the Commission of Seven on grade placement in arithmetic. 38th Yr. Book, Nat. Soc. Stud. Ed., Part I, Chap. 15, 299–324. *259, 260*

Watson, G. B. (1925) The measurement of fairmindedness. Contrib. to Ed., Teach. Col., Columbia Univ., No. 176. *486, 487, 488, 489, 490, 517, 596*

——. (1929) Orient and occident: an

opinion study. (Unpublished work on file at Inst. of Pacific Relations, 129 E. 52nd St., N. Y. City.) A digest of a part of this study appeared under the same title in Religious Educ., 21, 322–328. *515*

Watson, J. B. (1913) Behaviorism. The People's Institute Publishing Company, N. Y. Pp. 248. *151*

Webb, E. (1915) Character and intelligence. Brit. J. Psychol., Mon. Suppl. No. 3. *591*

Wechsler, D. (1931) The incidence and significance of finger nail biting in children. Psychoanalytic Rev., 18, 201–209. *658*

——. (1939) The measurement of adult intelligence. Williams and Wilkins, Baltimore. Pp. 229. *334*

Weiss, A. P. (1925) A theoretical basis of human behavior. R. G. Adams Co., Columbus, Ohio. Pp. 428. *151*

Wellman, Beth L. (1937) Mental growth from preschool to college. J. of Exp. Ed., 6, 127–138. *726*

Wells, F. L. (1908) On the variability of individual judgment. Essays Philos. and Psychol., (James), 509–550. *694*

——. (1912) The relation of practice to individual differences. Amer. J. Psychol., 23, 75–88. *623*

Wesley, E. B. (1932) Tests in social and political terms. Scribner, N. Y. *238*

West, M. (1935) Definition vocabulary. B. Dept. Ed. Res., Ontario Col. Ed., No. 4. Pp. 105. *201*

West, P. V. (1923) A critical study of the right-wrong method. J. Ed. Res., 8, 1–9. *130*

Wheat, L. B. (1931) Free associations to common words. Contrib. to Ed., Teach. Col., Columbia Univ., No. 498. *537*

——. (1932) Four spelling rules. Elem. Sch. J., 32, 697–706. *216*

Whipple, G. M. (1903) Studies in pitch discrimination. Amer. J. Psychol., 14, 553–573. *427*

Whipple, G. M. (1908) Vocabulary and word-building tests. Psychol. Rev., 15, 91–105. *298*

——. (1911) Manual of mental and physical tests. Vol. I and II. Warwick and York, Baltimore. Pp. 690. *300, 371, 372, 371*

——. (1919) Whipple's group tests for grammar grades. Public Sch. Pub. Co., Bloomington, Ill. *320*

Whipple, G. M., *et al.* (1920) The national intelligence tests. J. Ed. Res., 4, 16–31. *320*

Whisler, L. D. (1934) Multiple-factor analysis of generalized attitudes. J. Ed. Psychol., 5, 283–297. *507, 508, 588*

Whitford, W. G. (1920) Empirical study of pupil ability in public school art courses. Elem. Sch. J., 20, 1. *395*

Whitley, M. T. (1911) An empirical study of certain tests for individual differences. Arch. Psychol., No. 19. *623*

Whitman, E. C. (1925) A brief test series for manual dexterity. J. Ed. Psychol., 16, 118–123. *372*

Whorley, K. S. (1933) An experimental investigation of the sensitivity of children to compositional unity. Psychol. Monog., 45, No. 1, 26–45. *405*

Wiersma, E. (1906) De Sekundärfunktion bei Psychosen. J. Psychol. u. Neur., 8, 1–24. *589*

Wiley, L. E., and Wiley, A. M. (1937) Studies in the learning function. Psychometrika., 2, 1–20. *611, 612*

Wilke, W. H. (1934) An experimental comparison of the speech, the radio, and the printed page as propaganda devices. Arch. Psychol., No. 169. *512*

Wilkins, M. C. (1928) The effect of changed material on the ability to do formal syllogistic reasoning. Arch. Psychol., No. 102, 79. *25*

Willett, G. W. (1918) Permanence of pupil interests. Sch. and Soc., 9, 334–338. *476*

Williams, E. J. (1933) A technique for testing color harmony sensitivity in young children. Psychol. Monog., 16, 16–50. *401*

Williamson, E. G. (1937) Decreasing accuracy of scholastic predictions. J. Ed. Psychol., 28, 1–16. *257*

Willing, M. H. (1926) Individual diagnosis in written composition. J. Ed. Res., 13, 77–89. *212*

Wilson, G. M. (1923) Wilson language error test. World Book Co., Yonkers, N. Y. *212, 213*

Wingfield, A. H. (1928) Twins and orphans; the inheritance of intelligence. J. M. Dent, London. Pp. 127. *716, 722*

Witmer, L. (1894) Zur experimentellen Aethetik einfacher räumlicher Formverhältnisse. Phil. Stud., 9, 96–144, 209–263. *692*

——. (1911) Courses in psychology at the summer school of the University of Pennsylvania. Psychol. Clin., 4, 245–273. *344*

Wood, E. P. (1927) Improving the validity of collegiate achievement tests. J. of Ed. Psychol., 18, 18–25. *121*

Woodall, C. S. (1932) The children of mentally defective and mentally retarded mothers. Proc. and Addr. Am. Assn. Stud. Feeble-Mind., 36, 328–354. *715*

Woodrow, H. (1937) Absolute scaling applied to the data yielded by the method of constant stimuli. J. Exp. Psychol., 20, 1–28. *690*

——. (1938) The effect of practice on test intercorrelations. J. Ed. Psychol., 29, 268–278, and 561–572. *631*

——. (1938) The relation between abilities and improvement with practice. J. Ed. Psychol., 29, 215–230. *631*

——. (1939) Factors in improvement with practice. J. Psychol., 7, 55–70. *631*

Woods, F. A. (1906) Mental and moral heredity in royalty. Henry Holt, N. Y. Pp. 132. *715*

Woodworth, R. S. (1917) Personal data sheet. C. H. Stoelting Co., Chicago, Ill. *538*

——. (1938) Experimental psychology. Henry Holt, N. Y. Pp. 889. *364, 367, 578*

Woodworth, R. S., and Wells, F. L. (1911) Association tests. Psychol. Monog., Vol. 13, No. 5. *299, 348*

Woody, C. (1916) Measurement of some achievements in arithmetic. Teach. Col. Contrib. Ed., Teach. Col., Columbia Univ. *676*

Wrenn, Gilbert. (1934) A study habit inventory. Stanford Univ. Press. *246, 247*

Wrightstone, J. W. (1934) Civic beliefs and correlated intellectual and social forces. School Rev., 42, 53–58. *515*

——. (1936) Appraisal of experimental high school practices. Teach. Col., Columbia Univ. Pp. 194. *236, 239, 240*

——. (1938) Appraisal of newer elementary school practices. Teach. Col., Columbia Univ. Pp. 221. *187, 188*

Wyman, J. B. (1925) Interest tests of a group of gifted children. Genetic studies of Genius, Vol. I (ed. L. M. Terman). Stanford Univ. Press. Pp. 455. *537*

Yerkes, R. M., Bridges, J. W., and Hardwick, R. S. (1915) A point scale for measuring mental ability. Warwick and York, Baltimore. Pp. 218. *269, 282*

Yerkes, R. M., *et al.* (1921) Psychological examining in the United States army. Mem. Nat. Acad. of Sci., 15. U. S. Govt. Printing Office, Washington, D. C. *96, 307, 318, 319*

Yoakum, C. S., and Yerkes, R. M. (1920) The army mental tests. Henry Holt, N. Y. Pp. 294. *310*

Ziegler, W. A. (1934) Minn. rate of manipulation test. Ed. Test Bur., Minneapolis. *376*

GLOSSARY AND SUBJECT INDEX

GLOSSARY AND SUBJECT INDEX

Items in **boldface** are those for which definitions are given. Only those definitions are included which are emphasized in this book. For other usages one should consult Warren's (1934) *Dictionary of Psychology* or English's (1934) Student's *Dictionary of Psychological Terms.*

Ability: *A power, which can be demonstrated, to perform an act. It may be either inherited or acquired. Usually it is largely dependent upon training*, 440.

Absolute scaling, 675.

Abstraction: *Ability to consider a quality apart from a unified pattern or concrete object*, 190.

Academic: *Concerning a course of study which does not deal primarily with arts or practical skills* (see *Achievement tests*). Choices, 446; interests, 439.

Acceleration: *Rate of change, positive or negative*, 608.

Accomplishment, art, 403; indices of academic, 248; in music, 430.

Accomplishment quotient (AQ): *A ratio contrasting educational and mental development*, EA/MA or EQ/IQ, 250.

Achievement test: *A test for determining level of skill or range of information, either academic or vocational.* Cardinal objectives of education, 187; correlational analysis, 251; tests, 91, 187.

Acquired: *Characteristics which are gained from specific individual training, not from maturation alone* (see *Practice and learning*), 601.

Adjustment, conduct measures, 549; correlational analysis, 583; definitions, 521; economy of method, 579; evaluation of items, 594; free association, 523; general scales, 545; laboratory methods, 570; psychiatric classification, 540; psychological classification, 540; questionnaires, 538; records of observations, 563; reliability, 580; validity, 581.

Age equivalent: *The age score assigned to a given point score on tests provided with age norms*, 62.

Age level, assignment to, 282.

Age norms: *The achievement of subjects of various age groups, usually expressed as averages and dispersions*, 62.

Age score: *A score taken to represent the average of an age group*, 62.

Agencies using tests, 9.

Aggressiveness: *Carrying out one's own plans despite opposition.* Deception, 550; will, 590.

Agility tests, 367.

Alertness test: *Any test where speed and accuracy are demanded on many short mental tasks* (see *Rate tests*), 95.

Algebra tests, 229.

Alpha test: *U. S. Army battery of eight language tests designed to indicate general intelligence*, 64, 229.

Altitude test: *A test with various degrees of difficulty designed to show one's levels of mental ability*, 94.

Ambiguity, 122, 595.

Ament: *Mentally defective from birth to a marked degree. Synonymous with feebleminded*, 266.

Analogies test: *A reasoning test in which a relation between two or more items must be discovered and applied in selecting other items.* Spatial, 332; verbal, 306.

Analysis: *A logical division of a pattern into its constituent parts.* Factorial, 135; of a group of scores, 143; of a score, 141; of status, 156.

Anatomical age: 1. *Degree of physical development expressed in age norms.* 2. *Usually measured by X-rays of carpal bones*, 645.

Anthropometric measurement: *Physical measures of bodily dimensions*, 159.

Appraisal: *Any record or estimate of achievement or ability.* Individual, 7; varieties of, 15.

Aptitude test: 1. *A test which is designed to predict future success in some field.* 2. *A test which is not significantly affected by practice or training; hence, a measure of capacity*, 92.

Arithmetic mean: *A measure of central tendency; commonly called the average*, 50.

Arithmetic tests, 227.

Array: 1. *Any series of results arranged in a table.* 2. *In a correlation chart, a single row or column*, 42, 67.

Art, and intelligence, 409; analytical studies, 411; appreciation, 386; composition, 403, 406; drawing ability, 405; nature of, 383; preferences for pictures, 393; responses to color, 399; responses to lines, 397.

Articulation, 173.

Ascendance: *Leadership or domination*, 541, 571.

Assembly test: *One on which mechanical objects are to be put together to form a machine or part of a machine*, 28.

Association test: *Any test in which a new series of relations must be learned. Usually limited to the digit-symbol type of test* (see *Free association*), 30, 314, 348.

Assumed mean: *The mid-point of the class-interval which is taken as the point about which the correction is made in computing the arithmetic mean by the short method. Same as guessed mean*, 52.

Attention: 1. *The process of narrowing the range objects to which one responds.* 2. *The process of observing certain objects so that they become clearer.* Tests, 24.

Attenuation: *A reduction in the apparent relation between two variables due to errors in measurement. Theoretically corrected by Spearman*, 79.

Attitude: *An evaluation of policy.* Comparison of techniques of appraisal, 516; definitions, 485; factorial analysis, 506; group norms, 519; modification of, 510; needed research, 520; reasons for, 498; relation to information, 515; toward church, 36; typical scales, 486.

Audiometer, 6.

Auditory: *Concerned with hearing*, 5, 173.

Auditory span, 28.

Autistic: *Thinking which is largely controlled by wishes and not checked with reality*, 587.

Average: *Any measure of central tendency, but commonly used to designate the arithmetic mean*, 50.

Average deviation or AD (same as *Mean variation* and *Mean deviation*): *The sum of the deviations from the mean, ignoring signs, divided by the number of cases*, 55.

B-Score: *A Grade Score in which units represent school grades and tenths represent months of school year;* e.g., *6.7 = Sixth Grade seventh month*, 63.

Basal year: *In the Stanford-Binet test, the year level at which a person passes all tests and is assumed or proven to pass a lower level test*, 271.

Behavior: *Any complex pattern or adjustment in contrast to constituent physiological activities*, 3, 157.

Bellevue Scale, 334.

Beta test (U. S. Army): *A battery of eight non-verbal paper and pencil tests designed to measure general intelligence*, 309.

Bias: 1. *Same as prejudice.* 2. *In sampling, a selection of cases partially in accordance with some principle other than chance. Would result in systematic error*, 97.

Bi-modal: *A frequency distribution or polygon with two distinct modes.* Curve, 53.

Binet-type test, 264.
Biographies, 38, 442.
Bookkeeping test, 244.
Borderline: 1. *Not easily classified.* 2. *Defective, mentally just above the class of feebleminded*, 290.
Breadth test, 95.
Brown-Spearman formula: *One for indication, the probable reliability of which may be achieved by lengthening a test* n *times*, 101.
Business achievements, 241.

CA (Chronological Age), 268.
California preschool schedule, 169.
California Test of Mental Maturity, 323.
Capacity: *Extent of power to learn or to profit from training, limited by inheritance*, 703.
Carpal bones: *Wrist bones*, 160.
Case history, 32, 38, 567; of interests, 442.
CAVD Scale: *A battery of four types of tests* (*Verbal Completion, Arithmetic Problems, Vocabulary, Directions*). Developed to measure general mental ability, by Thorndike, 105, 325.
Centile (same as *Percentile*): *The divisions between each one hundredth part of a group which has been arranged in order of size*, 47, 61, 295.
Central tendency: *A score representing the middle of a group, mean, median, mode*, 50. Uses of, 54.
Chance: 1. *The probability that an event will occur in the light of one's knowledge of similar situations.* 2. *A variation in score from an unknown cause.* Correction for chance success, 123: relationships, 69.
Character: *One's moral attributes and ideals*, 549.
Cheating, 550.
Chemistry test, 26.
Childhood, early, 151.
Chronological Age, CA (same as *Life Age*): *Period between birth and a given date*, 268.
City environment, 725.
Class interval (i) (same as *Step*): *A group of scores which are a part of a larger distribution. The size of* i *is its range of scores. The frequency of* i *is the number of scores it contains*, 41.
Clerical aptitude tests, 19, 23, 33, 223.
Clerical test, 242.
Clinical: *Diagnosis and treatment of sick persons, or those referred because of failure*, 11.
Coefficient of alienation (k): *An index of lack of relation between two variables*, 84.
Coefficient of variation: *An indication of the size of a dispersion of a group relation to its central tendency*, 661.
Combining scores, 62.
Comparisons, group, 58; individual, 58, 59; qualitative, 4; quantitative, 4.
Completion test: *A test in which one must fill in a blank to complete a context by word or picture*, 112.
Composition, art, 403; English, 195; literature, 424; music, 443.
Comprehension, 199; general reading, 219; general vocabulary, 201; technical vocabulary, 206; words, 200.
Concept: *A general idea, which can be applied to several situations*, 190.
Conduct, 549
Consistency, internal, 102.
Consonance: *A pleasant blending of tones*, 428.
Constancy of IQ: *Shown by the amount of change in a person's intelligence quotient over a period of years*, 292
Constant Error (same as *Systematic Error*): *A type of error in appraisal which displaces the average of a group of scores in one direction*, 97.
Construction, of questionnaire items, 595; scale, 675; of test items, 116: of tests, 90.
Controlled-association test: *One in which a particular type of response is required for a stimulus*, 29.
Co-operation, 552.
Co-ordination: *The efficient timing and direction of forces which are just strong enough to accomplish a desired goal.* Motor, 21, 376.

Correction, for chance success, 123; for guessing, 127.

Correlation coefficient: *An index of relationship which ranges in value from plus one (+ 1.0) through zero to minus one (− 1.0). There are a number of indices of correlation.* Analysis of, 145; applications, 76; calculation, 67; errors, 77; interpretations, 86; probable errors, 83; product-moment, 71; rank order, 72.

Correlation matrix, 138, 145, 335.

Correlation table: *A two-way or double entry table which shows the relation between pairs of measures of the same individuals or items* (same as *Scatter Diagram*), 67.

Covariation, 89.

Criterion: *A standard by which the validity of a test may be determined.* Alpha test, 305; Binet tests, 281.

Critical age, 154.

Critical score: *A score which distinguishes between probable successful and unsuccessful candidates*, 53.

Cube Imitation Test, 349.

Cumulative frequency: *A frequency table in which each step shows the sum of all frequencies up to and including its own.* Curve, 45.

Curve, cumulative, 45; growth, 184, 665; normal frequency, 44; ogive, 45; practice, 605; regression, 78; shape, 54.

CV (see *Coefficient of Variation*), 661.

Cycloid: *Resembling cyclothymia*, 587.

Cyclothymia: *Marked alteration of mood between excitement and depression. In extreme form it becomes manic-depressive psychosis*, 587.

Deafness, 5, 6.

Deciles: *The dividing lines between tenths of a group of scores arranged in order of size*, 47.

Deduction: *A mode of reasoning which solves a problem by applying a premise or law to a particular situation*, 333.

Dementia praecox (nearly the same as *Schizophrenia*): *A psychosis characterized by a violent conflict of one's ideals. A functional splitting of memory and deterioration. There are four main clinical subdivisions: Simple, hebephrenic, catatonic, and paranoid*, 540.

Detroit Mechanical Aptitude Test, 30, 376.

Development, order of, 177.

Developmental Age: *Similar to Mental Age, but also includes indices of physical and social development*, 160.

Developmental zero: *Time at which an organism first exists as an organism*, 607, 630, 643.

Deviation: *The amount by which a score differs from central tendency of a group of measures*, 51.

Dexterity tests, 371.

Diagnostic test: *One designed to yield separate scores for independent and important skills.* Learning difficulties, 215; mental abilities, 337; reading, 218.

Differentiation: 1. *The process of cells whereby the cells in an organism develop various structures.* 2. *Roughly used to apply to any discrimination or classification*, 638.

Difficulty of an item: *Usually indicates the proportion of a group which successfully completes an item*, 279. Effects on obtained factors, 340.

Digit span tests, 24.

Digit symbol tests, 30, 314, 348.

Dimensions, of a group of scores, 50.

Discrimination: *The detection of differences of any sort.* Among adjacent age groups, 280; between bright and dull, 277; in literature, 415; in music, 426; sensory, 17.

Disorientation: *Loss of accurate knowledge of one's name and location*, 276.

Dispersion (same as *Scatter* and *Variability*): *The extent to which a group of measures differ from some point.* Changes due to growth, 659; changes due to practice, 623; of scores, 50; uses of measures of, 58.

Distribution of scores: *A table or chart which shows the frequency of each score or group of scores*, 40.

Drawing ability, composition and representation, 406; development of, 405.

Echolalia, 180.

Educational Age (EA): *Nearly synonymous with achievement age; however, it usually applies to a pupil's standing in a number of school subjects while achievement age frequently refers to a single subject*, 248.

Educational Quotient (EQ): *The Educational Age divided by the life age*, 248.

Educational tests, uses of, 10.

Emotion: *A more or less violent response to a disturbing situation* (see *Adjustment*), 521, 584, 590.

Empathy, 387.

English usage, 210.

Environment, changes in home, 719; special training, 724.

Epilepsy: *A nervous disease marked by convulsions and usually by loss of consciousness*, 540.

Equal-appearing-interval scale: *One in which judges agree that all the steps in a scale appear equal*, 695.

Equal mental ages, 287.

Equivalent forms: *Standard tests which yield scores approximately equal to each other, but which have different items or context.* Reliability, 100.

Error of grouping: *An error introduced by the practice of placing together in class-intervals unlike measures*, 78.

Errors, random, 79, 97; systematic, 97; which lower correlations, 78; which raise correlations, 81.

Essay test, 113, 195.

Examiner, requirements of, 12; training of, 13.

Exercise (same as *Practice*): *Attaining skill by repeated performance*, 639.

Experiment: *A procedure of arranging conditions so that observations of the causes of a particular phenomena can be made systematically*, 9.

Experiments in heredity, 717.

Expression, verbal, 190.

Extroversion: *A tendency to attend to and gain satisfaction from external stimuli*, 541.

Eye fixations, 655.

f.: *A symbol for the number of cases in a subdivision of a group*, 52.

Factor: 1. *Any constituent of a pattern*, 135. 2. *A mathematical factor is a parameter of a set of measurements*, 137. 3. *A psychological factor is any aspect of a pattern of behavior which is found to be unique in more than one individual*, 137. Chance factors, 145; common factors, 145; general factor, 139; group factor, 139; specific factors, 146.

Factorial analysis, advantages, 147; limitations, 146; of abilities, 337; of achievement, 251; of attitudes, 506; of motor skills, 378; of practice, 630; of traits, 588; results, 148; Thurstone's method, 141.

Fatigue, 140.

Feeble-minded, 266, 290, 531.

First quartile (Q_1): *The point on a scale of values such that 25% of the measures are less and 75% are greater*, 55.

Focal items, 154.

Following directions tests, 114, 299.

Fore exercise: *A preliminary or practice exercise for the purpose of giving the pupil experience with the particular test situation*, 322, 356.

Foreign languages, 240.

Forgetting, 634.

Formal tests, 91.

Formula: *An expression written with letters and digits to represent a natural law or usual relationship*, 52.

Free association tests, 30; pictures, 523; situations, 557; words, 536.

Frequency: *The number of measures in a given class interval of a tabulation*, 40.

Frequency curve: *A line representing a distribution of measures*, 41. Cumulative, 45; normal, 44.

Frequency polygon: *A representation of a frequency distribution by means of bars or areas*, 43.

Frequency table: *A table showing the number of cases in each class-interval of a distribution of scores*, 40.

Gaussian Curve (same as *Normal probability curve*), 44.

General factor (*g* factor), 139, 592.

Geometry tests, 229.

Grade equivalent: *The school grade which corresponds to a pupil's test score*, 62.

Grammar tests, 210.

Graphic rating scale: *A rating scale in which a check is made on a straight line to indicate degree of some trait*, 700.

Graphology: *The study of human characteristics as shown by handwriting*, 194.

Group, correlational analysis, 337; intelligence tests, 298; testing, 20.

Growth, correlational studies of children, 182; cycles, 671; dispersion of scores, 659; essential conditions for growth studies, 641; individual curves, 665; intellectual, 650; nature of, 638; practice and growth curves, 612; relationships among patterns, 663; scales, 672; theoretical considerations, 668; typical results, 643.

Guessing, 110; chance success, 123; told not to guess, 127.

Guessed mean (same as *Assumed mean*): *A score taken arbitrarily in calculating the arithmetic mean by a short method*, 52, 71.

Halo effect: *The tendency in many raters to let a general opinion about a person influence judgments of independent traits*, 704.

Handwriting, 193.

Heinis curve: *A growth curve proposed by Heinis*, 673.

Heredity: *The transmission of characteristics through germ plasm* (see *Native traits*), 713.

Hierarchy: *A graded order from low to high. A matrix is said to have a hierarchial order when there is a decrease in the size of coefficients from one corner towards the other three corners*, 138.

Histogram: *A chart in which the frequencies of class intervals in a distribution table are represented by rectangular areas*, 41.

Hypothesis: *A theory which is not conclusively proved, but is supported by some data*, 9.

Hysteria: *A psychoneurosis characterized by emotional disorders of various sorts*, 540.

i.: *The symbol for a class interval*, 52.

IQ (see *Intelligence Quotient*).

Identical twins: *Very similar twins, presumably formed from the division of a single fertilized ovum*, 716, 722.

Idiot: *The lowest division of mental ability. Often partially defined as those persons whose IQ's are below 25*, 290.

Imbecile: *The next to the lowest class of the mentally deficient. Often partially defined as having IQ's ranging from 25 to 49*, 290.

Incentive: *An external stimulus to which one responds with considerable force; opposite of an indifferent stimulus*, 441.

Index: *Any ratio of one dimension or parameter to another.* Accomplishment, 248.

Individual differences: *The observed or measured differences between individuals in capacity, ability, industry*, etc. Due to growth, 659; due to practice, 623; indicated by measures of dispersion, 50; native differences, 713.

Individual predictions, 73.

Individual test: *One which a single examiner can give to only one person at a time*, 20.

Individuation, 638.

Induction: *In logic the process of generalizing, that what is true of a few is also true of a larger group.* Tests, 25, 332.

Industrial, use of, tests, 10.

Infancy: *A period variously defined, often taken to include from birth to 2 or 3 yrs. Followed by early childhood to the 7th yr.*, 153.

Inference: *Any process of concluding that a statement is true because it is based upon other statements which are known to be true.* Test, 25, 115, 332, 488.

Inflection point, 44.

Informal test: *A test distinguished from a standardized test by little standardization and only local usage*, 91.

Information test: *One in which success depends primarily upon recall or recognition of facts*, 29.
Inherited (same as *Native*, *Innate*, *Inborn*): *A characteristic which is passed on from one generation to others through the germ plasm*, 713.
Inhibition, 555.
Insane: *A legal term referring to any person who suffers from a severe mental disorder*, 540.
Instinct: *An* inherited *tendency to perform a complicated act when a particular combination of stimuli and maturity is experienced*, 639.
Instruction, optimum age for, 259.
Integration, 523, 559, 575.
Intelligence: 1. *The power to learn, or degree of plasticity.* 2. *The power to solve problems.* 3. *The inborn capacity to develop skills in abstract thinking.* 4. *Thorndike suggested three: social, mechanical, and abstract intelligence*, 264, 298.
Intelligence quotient, IQ: *An index of relative brightness or rate of development. The Mental Age (MA) divided by the Chronological Age (CA).* Constancy of, 292; for adults, 291; interpretation of, 288; similarity: of average, 289; of standard deviations, 289; substitutes, 295.
Intelligence test: *Any group or individual test which is designed to measure general mental ability or intelligence.* Group, 298; individual, 264.
Interest (same as *Desire*, *Wish*, *Like*): *A readiness to respond to a set of stimuli.* Case histories, 442; definitions, 439; factor analysis, 470; inventories, vocational, 458; job satisfaction, 449; needed research, 483; permanence, 473; prediction of scholastic success, 477; prediction of vocational success, 481; reasons for academic choices, 446; reasons for vocational choices, 447.
Internal consistency, 102, 122.
Interpolation: *The process of locating an intermediate point between two known points in accordance with a rule*, 47.
Interpretation: *An explanation of results by showing their causes or their significance.* Of intelligence quotients, 288; of mental ages, 285; of pictures, 533; of scores, 39.
Interviews, 37, 570.
Introversion: *A tendency to attend to and gain much satisfaction from one's own thoughts*, 541.
Inventories: *A check list of abilities and personal characteristics.* Comparison of, 466; of preferences, 457.

Job satisfaction, 449.
Judgment: *A process of deciding between two or more choices.* Advantages of various forms, 701; combinations of, 707; halos and logical errors, 705; methods of measuring, 35, 691; paired comparison, 692; rank order, 694; rating scale, 699; validity of, 708.

Kent-Rosanoff test: *A free association test of 100 words*, 30, 536.
Knox cube test: *A non-verbal test of immediate memory span in which the task is to repeat a series of taps on four cubes*, 344.
Kurtosis (Ku): *The flatness of a frequency curve compared to the normal curve*, 57.

Language, foreign, 240: measures, 189; order of development, 179; techniques of reading, 209.
Laterality, 174.
Law of filial regression, 68.
Learning: *A general term for any process which changes an animal's mode of behavior through experience.* Measures, 32.
Letter grades, 47.
Lie detecting, 550, 578.
Life Age (same as *Chronological Age*): *Time elapsed since birth*, 268.
Literature, appreciation of sounds, 419; composition, 424; discrimination of style, 415; information, 423; measurement in, 415; merit of prose and poetry, 420.
Logic: *A science of classifying formal arguments into types*, 25.

Logical errors, 704.
Logs, 37, 562.
Longitudinal studies: *Those which collect data on an individual over long periods of time*, 665.

MA (see *Mental Age*).
MD (see *Average Deviation*).
Man-to-man rating scale: *Made of one's own acquaintances*, 701.
Mania: *A violent conduct in which the subject is somewhat disoriented*, 540.
Manic-depressive psychosis: *A severe disorder characterized by great excitement followed by a stuporous condition*, 540.
Manual training test, 355.
Matching test: *The task here is to select an item from one group which belongs with an item in another group*, 112. Mechanical, 356.
Mathematics, 227; arithmetic tests, 227; algebra, geometry, and trigonometry tests, 229.
Maturation: *Growth of an organism due to innate mechanisms in a normal environment — as distinct from learning*, 639.
Maturity, age of, 291.
Maze test, 351, 353.
Mean: *Usually signifies the arithmetic mean, in which is the point on the scale about which the sum of the deviations is zero*, 50, 54. Changes in, 614.
Mean variation, MV (see *Average Deviation*).
Measurement: *A quantitative comparison of an individual with a standard scale*, 4.
Mechanical tests: 343. Assembly for girls, 361; Minnesota assembly, 28, 358; paper and pencil, 354.
Md (see *Median*).
Median (Md): *The midmost score in a group arranged according to size*, 50, 53. Scale value, 518.
Memory: *An ability to retain one's experiences* (same as *Retention*), 28.
Memory span tests, 24.
Mental: *Usually refers to conscious behavior as distinct from physiological and involuntary acts*, 4.
Mental ability: *Used somewhat synonymously with intelligence. Theoretically, a general power to learn, remember, and solve problems*, 10, 264, 298.
Mental age: *The average score on a particular mental test of persons in a small age group. Less inclusive than development age.* Changes in meaning, 285; interpretation, 285; of national groups, 319; scaling, 676; substitutes for, 295.
Mental deficiency: *Any serious lack of mental ability, including borderline, moron, imbecile, and idiot*, 290.
Mental units (MU), 274
Merrill-Palmer Scale, 162.
Metropolitan Readiness Tests, 171.
Michigan Aiming Test, 18.
Michigan Non-verbal Series, 353.
Michigan Number Comparison Test, 19.
Michigan Speed of Reading Test, 49.
Michigan Vocabulary, Profile, 206.
Mid-point: *The exact middle of a step or class interval in a frequency table*, 51.
Minnesota Mechanical Assembly Test, 28.
Minnesota Paper Form Board, 357.
Minnesota Preschool Test, 165.
Minnesota Rate of Manipulation Test, 21.
Minnesota Spatial Relations, 359.
Minnesota Vocational Test for Clerical Workers, 23.
Mode: *In a frequency table, the score or class interval which has the highest frequency* (see *Bimodal*), 50, 52.
Moral knowledge, 489, 556.
Moron: *The highest class of feebleminded individuals. Often partly defined as those with IQ's between 50 and 75*, 290.
Motivation: *A general term for any combination of needs and incentives*, 439, 603.
Motor, ability: *The power to perform complex co-ordinated movements; includes strength and endurance*, 21. Co-ordination, 363; co-ordination and spatial thinking, 647; correlational analysis, 378; rhythms, 374; tests, 376.
Multiple choice test: *Items which require a person to select one or more correct*

responses out of a group of several possible responses, 110.
Music, 425; appreciation, 428; correlational analysis, 433; criticism of tests, 436; discrimination of elements, 426; information, 430; profile, 435; racial differences, 437.

N or **n:** *A symbol for the number of cases in a group*, 52.
Native (same as *Hereditary*), correlational studies, 715; effects of environment, 719; experimental studies, 717; measurement of differences, 713; selective breeding, 713; special training, 724.
Negative correlation: *A relationship in which the larger values of one variable correspond to the smaller values of the other variable*, 69. Among personality traits, 586.
Negativism: *A tendency to do the opposite of what is suggested*, 563.
Neurasthenia: *A condition characterized by real or imagined lack of energy*, 540.
Neurotic tendencies, 541.
Non-linear relationship: *A regression line which is markedly curved*, 70, 78.
Non-verbal tests, 344.
Norm: *A standard pattern for a group or an individual* (see *Profiles*, *Scales*), 50.
Normal probability curve (same as *Gaussian Curve*): *A bell-shaped distribution curve found very widely in biological data, and used as a standard*, 44, 289, 692.
Number series completion test: *One in which omissions in logical arrangements of numbers are to be filled in by the application of a rule which must be discovered from the arrangement*, 304.

Objective: *A term used in describing tests in which there is small disagreement among experts as to correctness of response*, 98, 110.
Objectivity: *Not subject to systematic errors*, 97.
Observational method: *Uses direct attention upon the behavior to be appraised, as contrasted with a rating from memory*, 562.
Ogive: *A cumulative frequency curve*, 45.
Omnibus test: *One composed of a variety of tasks usually appearing in random order*, 322.
Opinion: *A judgment concerning the truth of a statement*, 486, 558.
Opposites test: *A test calling for the opposite of each of a list of words or pictures*, 305.
Oral reading test: *One which requires passages to be read aloud*, 192.
Organization, paragraph, 224.
Orthogonal: *At right angles. When used of regression lines, it means that the variables are unrelated*, 143.
Overlapping: *The extent to which scores in two arrays fall within the same range*, 59.

PE (see *Probable Error*).
Painting, 406.
Paired Comparisons Method, 692.
Parameter: *Any well-defined dimension of a group or part of a group of scores, such as a mean, S.D., or mathematical factor*, 141.
Paranoia: *A psychosis marked by well systematized delusions of persecution and grandeur with little deterioration*, 530, 540.
Paranoid: *Resembling paranoia*, 530, 540.
Partial correlation coefficient: *An index showing the relation of two variables with the effect of other measures of variables removed statistically*, 82.
Percentage of determination, 87.
Percentile: *The scores which divide the total number of cases in a given frequency distribution into 100 equal parts*, 55.
Per cents of the average (PA), 274.
Perception, 22; of mechanical detail, 383; tests of, 223.
Performance scales: *Scales designed to measure mental ability by tests of manipulation and assembly, usually without language*, 344.
Perseveration: 1. *A tendency of an idea to return again and again.* 2. *Inability to shift quickly from one type of response to another*, 140, 589.
Persistence, 555.

Personal inventory: *A list of items descriptive of human adjustments*, 549.
Personality (same as *Self*): *The structures and responses which characterize a person*, 545.
Physical sciences, 234.
Physiological age: *Similar to developmental age, but based upon the average development of bodily structures. Often inferred from ossification of wrist bones*, 160.
Physiological measures, 16.
Pitch: *The high or low character of a tone determined principally by its frequency of vibration*, 426.
Plateaus in practice curves, 626.
Poetry, 420.
Point-hour-ratio (same as *Credit-hour-ratio*): *The number of points or credits earned in a course of study divided by the number of hours of class work*, 258.
Point scale: *A numerical series of scores based directly on the number of items completed successfully*, 305, 675.
Political issues, tests of, 11, 486.
Postural Behavior, 157.
Power test: *One which measures complexity of achievement by allowing enough time for each person to attempt all the levels of difficulty in which he can possibly succeed*, 95.
Practice effect: *The effect which previous experience with a test has on a later trial.* Avoidance of, 633; changes in group dispersions, 623; conditions for measuring, 602; correlation of trials, 625; curves, 607; definitions, 501; factorial analysis, 631; measured results of, 613.
Precision: *Fineness of movement*, 17 (see *Steadiness*, 373).
Prediction, 63; limitations of, 14; of practice effect, 625; of probable scores, 65; scholastic, 256; vocational, 261.
Preference tests, artistic, 35; music, 428; prose and poetry, 420.
Prehension, 177.
Prejudice (same as *Bias*): *Judgment on the basis of some previous experience rather than the present*, 487.
Pre-school: *Education during the third or fourth year prior to the kindergarten or the first grade*, 165, 168.
Primary traits, mental, 329; temperamental, 586.
Probability: *The likelihood that an event will occur, estimated from previous similar situations. Often stated as chances per hundred* (see *Standard error*), 74.
Probable error (PE): *An index of dispersion of a theoretical normal distribution, which indicates ½ the range of the middle 50% of cases*, 55. Of correlation coefficient, 83.
Product-moment correlation coefficient: *One secured by comparing the sum of the products of the deviations of two variables with minimum possible sum*, 71.
Profile chart: *A chart which shows one's place in a group on several different appraisals*, 61, 158. Music, 425; vocational interests, 467.
Prognostic test (similar to *Aptitude test*): *A test designed for the prediction of future achievement based upon present performance*, 92.
Progressive education: *Any of a group of schools which have placed great emphasis upon well-rounded personal development in a co-operative society*, 188.
Propaganda: *An indirect effort to make others adopt some belief*, 510.
Prose, 420.
Psychiatry: *The practice of curing mental disease*, 540.
Psycho-galvanic reflex: *A change in the electrical polarity of a portion of the body in response to a stimulus*, 367.
Psychological tests, 94.
Psychophysical measures, 17.
Puberty: *Period in early adolescence during which the sex organs become capable of functioning*, 671.

Q: *Semi-interquartile range, or quartile deviation*, 55.
Qualitative, comparisons, 4; validity, 97.
Quality scales: *A set of standard specimens for comparison with a given speci-*

men. Handwriting, 195; drawing, 406; English composition, 197.
Quartile: *The scores which divide a group into quarters*, 47.
Quartile Deviation (Q): *Half the range of the middle 50% of a group of scores*, 55.
Questionnaire: *A set of questions concerning a subject*, 538. Attitude, 451, 487; interest, 461; job satisfaction, 453.

Random errors, 79, 97.
Range: *Difference between the highest and lowest scores of a group.* Total, 55.
Rank: *Position assigned to a score in a series arranged in order of size*, 47. Order comparison, 694; order correlation, 72.
Rate tests: *Measures of achievement in terms of the number of tasks of uniform difficulty which can be performed in a specified time*, 95.
Rating: *An estimate of qualities or abilities*, 35. Scale methods, 699; rules, 707; validity, 708.
Raw score: *A numerical indication of achievement based directly upon the number of items correctly or incorrectly completed*, 39.
Reaction time: *The interval between the application of a stimulus and the subject's response*, 364.
Reading readiness, 170.
Reading tests, 219, 617.
Rearrangement test, 112.
Reasoning: *A form of thinking which involves conclusions from premises. In animal work reasoning is thought to be demonstrated when two temporally separate experiences are combined to solve a problem.* Tests, 24, 324, 333.
Recall test: *A test which demands that a context, word, or picture be completed from memory, illustrated by Completion or Essay type tests*, 30.
Recognition test: *One of the true-false or multiple-answer type, which exercises do not force the examinee to formulate the answer*, 29.
Regression equation: *The equation of the line which best fits the means of each class interval in a correlation table*, 73.
Relationships, of items, 77; of persons, 77; of scores, 77.
Relatives, of eminent men, 714; related pairs, 716.
Reliability: *The extent to which a score, or a central tendency, or any index remains consistent with itself on various occasions.* Coefficient, 100; of larger samples, 101.
Rho (ρ): *Correlation coefficient calculated from rank differences*, 72.
Rhythm, mental, 376; motor, 374; musical, 432.
Rorschach tests, 524; and mental age, 531; norms, 528; reliability, 527; scoring, 525; typical reports, 529.
Rural environment, 725.

SD (see *Standard Deviation*).
Sample: *Part of a larger group chosen to represent the whole.* Methods, 99; time sampling, 34, 562.
Sampling error: *The difference between any aspect of the sample and the same aspect of the whole group*, 34, 79.
Scatter: *The amount of variation in a group of scores* (same as *Dispersion*), 50.
Scale: *Any order by which individuals can be compared*, 91. Age, 282; graphic, 200; growth, 673; SD, 677; zero, 686, 689.
Scatter diagram (see *Correlation table*), 66, 70.
Scholastic predictions, 256.
Science tests, physical, 234; social, 236.
Score: *A numerical indication of the performance of a subject.* Dispersion of scores, 659; interpretation, 41; limits, 40; range of scores, 41; raw, 39; standard, 59, 61; standard deviation, 59; weighted, 63.
Scoring devices, automatic, 132; machine, 134; stencil, 131.
Seguin Form Board, 346.
Semi-interquartile range, Q (same as *Quartile deviation*): *One half the range of the middle 50% of a group*, 55.
Sensation: *The awareness that sense receptors are being stimulated*, 17.

Siblings: *Full brothers and sisters*, 716.
Σ (Capital Greek s, called *sigma*): *Symbol for* the sum of, 52.
σ (Small Greek s, called *sigma*): *Symbol for* Standard deviation, 52.
Single variable tests, 17, 19.
Situations, partly controlled, 20; slightly controlled, 34; well-controlled, 15.
Skewness: *A lack of symmetry in a frequency curve*, 54, 55, 57.
Social sciences, 236.
Spatial, motor co-ordination and spatial thinking, 647; visualization, 383.
Spearman-Brown Prophecy Formula: *A formula to determine the probable reliability of a test which cannot easily be repeated*, 101.
Specific factor (s-Factor), 146.
Speech tests, 192.
Speed tests, 95.
Spelling, 216.
Spot-Pattern Test, 24.
Standard Deviation (SD, or σ): *A measure of variability of a group*, 52, 55, 56, 71. Calculation, 58; formula, 67; of a score, 59; scaling, 676.
Standard error of estimate: *The probable dispersion of score in one variable estimated from one's score in another variable*, 74.
Standard error of an obtained score: *The probable variation of a score based on a reliability coefficient*, 75, 76.
Standard score: *A score which indicates one's position in a group by fixing the mean at 50 and allowing ten points for each standard deviation*, 60, 295.
Stanford Achievement Test, 63.
Stanford-Binet Test: *One of the revisions of the Binet Test made by L. M. Terman and his associates at Stanford University*, 270.
Steps, equivalence of, 105.
Study Habits, 246.
Subjectivity: *The degree to which measurement results are affected by personal factors or judgments* (see *Systematic errors*), 97.
Substitution test: *A test in which one symbol is substituted for another according to a code* (see *Association test*), 30, 314.
Suggestible: *One who is influenced more easily than the average*, 549.
Syllogism test, 25.
Symbol: *Any number, letter or form used to represent something else*, 42, 52.
Systematic error: *An error due to a bias or rule, which affects the average of a group of scores* (same as *Constant error*), 97.

T scale (same as the *Standard Score Scale*): Originally applied to a large group of 12-year-olds by McCall, 683.
Tabulation: *The process of grouping and recording scores according to size and finding the frequencies of the groups*, 40.
Temperament: *A fairly constant pattern of emotional adjustment*, 522, 540.
Tendencies, central, 50; group, 8.
Test (same as *Examination*): 1. *A situation in which a record of performance is secured directly, not by estimate.* 2. *Any set of questions which are to be answered during an examination.* Achievement, 91; agencies using, 9; altitude, 94; aptitude, 92; breadth, 94; formal and informal, 91; psychological, 94; single variable, 17; speed, 94; techniques, 13.
Test items, analysis, 121, 594; chance factors, 123; construction, 116; types, 109.
Testing, group, 20; individual, 20; limitations of, 13.
Time allowance, 115.
Time-sampling, 34; pre-school children, 176, 569.
Trade test: *One of knowledge or skill in a trade, usually a mechanical trade*, 354.
Trait, 522, 658.
True-false test: *A form of test item in which one simply indicates whether he thinks a statement is true or false*, 25, 111.
True measure: *A hypothetical score which represents the facts exactly. All known scores are considered to be approximations of true measures*, 75.
Twins, 722.

Unimodal: *A frequency distribution with but one mode*, 53.

Validity: *The degree to which a measuring instrument measures the pattern it purports to measure*, 97. Coefficient, 99.
Values, study of, 501.
Variability: *The amount of dispersion of a group of scores* (see *Range*, *Quartile Deviation*, *Standard Deviation*), 55.
Variable: *Any quantity which may exist in different amounts. Commonly used of tests and ratings*, 4.
Variance: *The square of the standard deviation*, 144.
Verbal relations, 383.
Vocabulary test: *A list of words to be defined*, 202.
Vocation: *An occupation or profession.* Choices, 447; factorial analysis, 470; interest inventories, 458; permanence, 473; predictions from interests, 481; predictions from tests, 261; satisfactions, 449.

W-factor: *The Will factor of Spearman and Webb*, 591.
Weighted scores, 63.
Work-limit: *An examination procedure in which each person completes the same tasks and the score is the time required*, 604.

x: *The symbol for a deviation from the mean*, 52.
X: *The symbol for a raw score*, 59.
x^1: *The symbol for a deviation from the guessed mean*, 52.

Y: *The symbol for a raw score*, 67.
y^1: *The symbol for the deviation from the guessed mean*, 67.
Youth-or-Inexperience factor, 383.

Zero: *A point taken as the origin of a scale*, 607, 630, 643, 686.

31047 0

SCIENCE ROOM

Date Due

Due	Returned	Due	Returned
NOV 19'63 ML	NOV 19'63		

DEC

151.2
G799m
c.2

Measurements of human behavior main
151.2G799m C.2
3 1262 03168 3039

KEEP CARD IN POCKET

ResLib

IT IS IMPORTANT THAT
CARD BE KEPT IN POCKET